VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky

Bakalářská práce

2017 Petr Blaha

VŠB – Technická univerzita Ostrava

Fakulta elektrotechniky a informatiky

Katedra kybernetiky a biomedicínského inženýrství

Hodnocení kvality řečových signálů na bázi virtuální instrumentace

Evaluation of Speech Quality Based on Virtual Instrumentation

2017 Petr Blaha

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a biomedicínského inženýrství

Zadání bakalářské práce

Student: **Petr Blaha**

Studijní program: B2649 Elektrotechnika

Studijní obor: 2612R041 Řídicí a informační systémy

Téma: Hodnocení kvality řečových signálů na bázi virtuální instrumentace
Evaluation of Speech Quality Based on Virtual Instrumentation

Jazyk vypracování: čeština

Zásady pro vypracování:

Bakalářská práce je zaměřena na hodnocení kvality řečových signálů na bázi virtuální instrumentace (LabVIEW). Cílem absolventské práce je návrh a softwarová realizace aplikace pro subjektivní i objektivní metody hodnocení kvality řečových signálů. Funkčnost navržené aplikace bude ověřena pomocí syntetických i reálných experimentálních dat.

Body zadání:
1. Literární rešerše problematiky subjektivního a objektivního hodnocení kvality řečových signálů.
2. Klasifikace a matematický popis metod pro hodnocení kvality řečových signálů (např. SNR, SSNR, GSNR, SSNRA, LLR, CD, PESQ, apod.).
3. Návrh a realizace aplikace pro subjektivní (časová, frekvenční, časově-frekvenční oblast) i objektivního hodnocení kvality řečových signálů ve vývojovém prostředí LabVIEW.
4. Tvorba experimentálních syntetických dat s definovanými parametry (SNR, apod.) pro účely testování metod hodnocení kvality řečových signálů.
5. Tvorba reálných dat pro experimenty pomocí měřicích mikrofonů.
6. Ověření funkčnosti vytvořené aplikace z pohledu adaptivního zpracování řečových signálů (určení kvality filtrace) pomocí NI LabVIEW Adaptive Filter Toolkit.
7. Diskuze dosažených výsledků.

Seznam doporučené odborné literatury:

[1] WITTASSEK, Tomáš. *Virtuální instrumentace I: učební text* [CD-ROM]. Ostrava: Vysoká škola báňská - Technická univerzita Ostrava, 2014. ISBN 978-80-248-3395-8.
[2] SOVKA, Pavel a Petr POLLÁK. *Vybrané metody číslicového zpracování signálů.* Vyd. 2. přeprac. Praha: Vydavatelství ČVUT, 2003. ISBN 80-01-02821-6.
[3] UHLÍŘ, Jan. *Technologie hlasových komunikací.* Praha: Nakladatelství ČVUT, 2007. ISBN 978-80-01-03888-8.
[4] BENESTY, Jacob, M. Mohan SONDHI and Yiteng HUANG (eds.) *Springer handbook of speech processing.* Berlin,Heidelberg: Springer-Verlag, 2008. p.XXXVI, 1176. ISBN 978-3-540-49125-5.
[5] MARTINEK, Radek, et al. A robust approach for acoustic noise suppression in speech using ANFIS. *Journal of Electrical Engineering.* vol.66, no.6, 2015, p.301-310. DOI: 10.2478/jee-2015-0050, Print ISSN 1335-3632, On-line ISSN 1339-309X.
[6] MARTINEK, Radek a Jan Žídek. Use of adaptive filtering for noise reduction in communications systems. In: *2010 International Conference on Applied Electronics (AE) 2010.* Plzeň: Západočeská univerzita, 2010. ISBN 978-80-7043-865-7, ISSN 1803-7232.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Ing. Radek Martinek, Ph.D.**

Datum zadání: 01.09.2016

Datum odevzdání: 28.04.2017

doc. Ing. Jiří Koziorek, Ph.D.
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlášení

„Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.“

V Ostravě dne 28. 4. 2017

Blaha
.....
Podpis

Poděkování

Rád bych poděkoval vedoucímu bakalářské práce doc. Ing. Radkovi Martinkovi, Ph.D. za odbornou pomoc a profesionální přístup při vedení bakalářské práce a Ing. Josefu Kročilovi za konzultace při vytváření této práce.

Abstrakt

V rámci této bakalářské práce byl vytvořen program pro objektivní a subjektivní hodnocení kvality řeči. Tento program byl vytvořen v programovacím a vývojovém prostředí LabVIEW. Pomocí tohoto prostředí bylo zpracováno množství nahrávek řeči a různých hluků jak ve venkovním prostředí, tak i z domácnosti. Jako objektivní metody hodnocení kvality řeči byly vybrány základní metody globálního odstupu signálu od šumu GSNR a segmentální odstup signálu od šumu SSNR s využitím intenzitního detektoru řečové aktivity. Jako subjektivní metoda hodnocení řeči byla vybrána metoda borcení časové osy DTW, přičemž byly zvoleny koeficienty PARCOR, jakožto příznakové vektory. Pro odfiltrování zašuměných nahrávek bylo použito adaptivní filtrace, algoritmů LMS a RLS. V práci jsou popsány výše zmíněné objektivní a subjektivní metody hodnocení kvality řeči.

Klíčová slova

Odstup signálu od šumu, detektor řečové aktivity, borcení časové osy, Částečné korelační koeficienty, adaptivní filtrace, metoda nejmenších čtverců, Rekurzivní metoda nejmenších čtverců

Abstract

In this bachelor thesis, a program for objective and subjective evaluation of speech quality was created. This program was created in the programming and development environment LabVIEW. Using this environment, a lot of speech recordings and various indoor and outdoor noises were processed. As the objective methods of speech quality assessment, the basic method of global signal-to-noise ratio (GSNR) and the segmental signal-to-noise ratio (SSNR) with an intensive voice activity detector (VAD) were chosen. As a subjective speech evaluation method, the Dynamic time warping (DTW) method was selected, with PARCOR coefficients being chosen as symptom vectors. To filter out smudged recordings, adaptive filtering, LMS and RLS algorithms were used. The previously mentioned objective and subjective methods of speech quality assessment are described in the thesis.

Key Words

Signal to Noise Ratio, Voice Activity Detector, Dynamic Time Warping (DTW), Partial Correlation Coefficients, Adaptive Filter, Least Mean Squares, Recursive Least Squares

Obsah

Seznam použitých symbolů a zkratek

$VoIP$	Komunikace přes IP (Voice over Internet Protocol)
SNR	Odstup signálu od šumu (Signal to Noise Ration)
$GSNR$	Globální odstup signálu od šumu (Global Signal to Noise Ration)
$SSNR$	Segmentální odstup signálu od šumu (Segmental Signal to Noise Ration
VAD	Detektor řečové aktivity (Voice Activity Detector)
DTW	Dynamické borcení časové osy (Dynamic Time Warping)
LMS	Stochasticky gradientní adaptace (Least Mean Squares)
RLS	Rekurzivní optimální adaptace (Recursive Least Square)

$\mathbf{A}$	Obraz testovaného slova
$\mathbf{a}(n)$	Vektor testovaného slova
$\mathbf{B}$	Obraz referenčního slova
$\mathbf{b}(n)$	Vektor referenčního slova
$D(\mathbf{A}, \mathbf{B})$	Celková vzdálenost mezi obrazy
E_n	Krátkodobá energie signálu
$e(n)$	Chybový signál adaptivních filtrů
$\mathbf{\Phi}_{kk}(n)$	Inverzní autokorelační matice
f_{vz}	Vzorkovací frekvence
G	Koeficient zesílení
$g(n, m)$	Funkce lokálního omezení DTW
$H(z)$	Přenosová funkce modelu
K_{VAD}	Počet segmentů s řečovou aktivitou
k_i	Koeficienty odrazu (PARCOR)
$\mathbf{k}(n)$	Ziskový vektor
λ	Faktor zapomínání
L	Počet analyzovaných segmentů
l	Délka signálu
μ	Konvergenční konstanta

M_n	Krátkodobá intenzita
M_p	Prahová hodnota intenzity
M_d	Úroveň hluku pozadí
N	Délka segmentu
$noise(n)$	Signál hluku
P_n	Výkon signálu šumu
P_s	Výkon řečového signálu
p	Parametr intenzitního detektoru
Q	Řád lineární predikce
$R_n(m)$	Krátkodobá autokorelační funkce
$s(n)$	Řečový signál bez hluku
σ_n^2	Rozptyl signálu šumu
σ_s^2	Rozptyl řečového signálu
$vad(n)$	Informace o řečové aktivitě
$w(k)$	Váhovací okénko
$\mathbf{w}(n)$	Vektor koeficientů adaptivního filtru

Seznam obrázků

Seznam tabulek

Úvod

Bakalářská práce se věnuje subjektivním a objektivním metodám pro hodnocení kvality řeči. Zpracování řeči, konkrétně počítačová analýza, nachází uplatnění v rámci řešení široké škály problémů souvisejících s komunikačními technologiemi. Jedná se zejména o detekci přítomnosti řeči v hlučném prostředí nebo o rozeznání řeči a převedení na text, což by například nevidomým osobám umožnilo a usnadnilo tvorbu dokumentů. Hlasové ovládání domácích spotřebičů není ještě v současnosti úplným standardem, nicméně je stále vyvíjeno a zdokonalováno.

Bakalářská práce se skládá z teoretické a praktické části. První kapitola popisuje kvalitu řeči. Druhá kapitola se zabývá popisem zpracování řeči a adaptivních filtrů. Poté je popsána lineární prediktivní analýza a Levinson – Durbinův algoritmus. Ve třetí kapitole jsou popsány subjektivní metody hodnocení kvality řeči, a metoda dynamického borcení času, pomocí které budou rozeznávány nahrávky a ohodnocena kvalita filtrace pomocí adaptivních filtrů. Ve druhé části třetí kapitoly jsou popsány objektivní metody hodnocení kvality řeči, metody odstupu signálu od šumu a intenzitní detektor řeči.

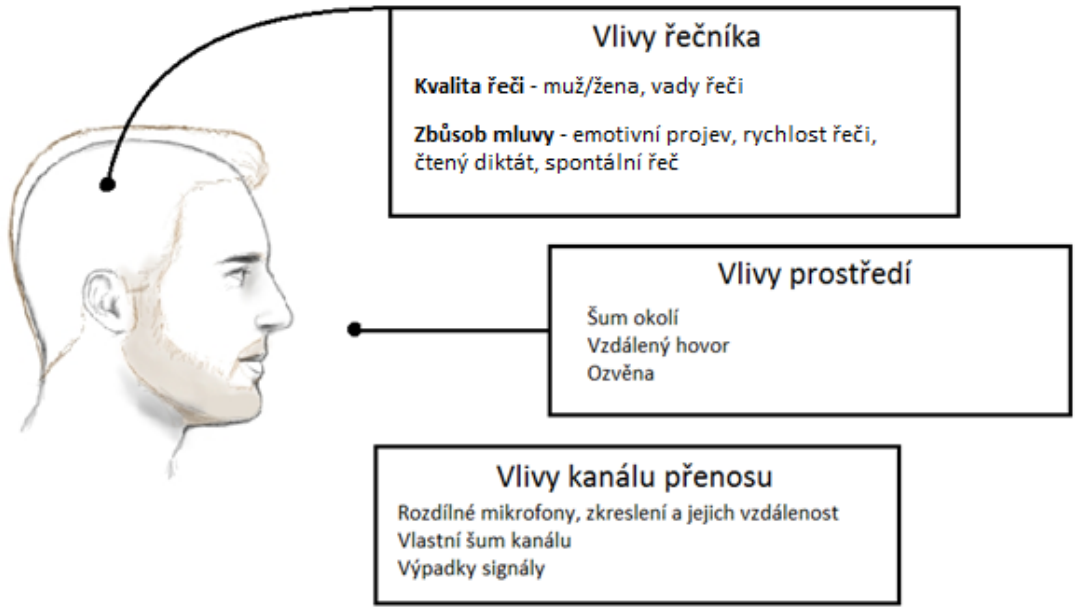
Čtvrtá kapitola shrnuje popis praktické realizace systému pro výpočty globálního a segmentálního odstupu signálu od šumu, detektoru řeči a výpočet vzdáleností mezi zvolenými slovy, navrženého ve vývojovém a programovacím prostředí LabVIEW. Praktická část obsahuje popis programu v prostředí LabVIEW a nachází se zde shrnutí výsledků.

Cílem bakalářské práce je z nastudovaných metod zpracovat program, který bude sloužit pro nahrávání vlastních řečových nahrávek a nahrávek hluku, ze kterých se bude pomocí objektivních a subjektivních metod posuzovat kvalita řeči. Pro filtraci zašuměných nahrávek budou používány adaptivní filtry, konkrétně LMS a RLS algoritmy, na kterých uvidíme jejich spolehlivost a kvalitu filtrace.

1 Posuzování kvality řeči

Řeč je nejstarším způsobem komunikace u lidí. Až postupem času se lidé naučili písmo a dokázali své myšlenky zaznamenávat i na papír. Řeč je u lidí plně automatická, a tak během ní můžeme vykonávat i další činnosti jako například psaní textu či chůzi.

Řeč může být posuzována na základě dvou aspektů. Jedním z nich je vnímání celkové kvality řeči a srozumitelnost. Celkové vnímání kvality je dojem posluchače, jak kvalitní je projev, přičemž výběr kritérií pro hodnocení kvality je ponechán na daném posluchači. Nicméně, jelikož slyšíme každý den přirozenou řeč při komunikaci mezi lidmi, můžeme si vytvořit tzv. referenční bod na stupnici kvality. Posluchači hodnotí řeč vzhledem k této referenci. Na druhou stranu srozumitelnost řeči definuje, do jaké míry je posluchač schopen rozlišit užitečnou informaci od okolního hluku, tzn. porozumět tomu, co slyší. Je udávána jako procentuální podíl správně identifikovaných reakcí relativního počtu odpovědí. Jako zkušební jednotky můžeme použít telefon, slabiky, slova nebo věty. Poslední dvě jmenované jednotky musí být jazykově smysluplné pro správné vyhodnocení. Vztah mezi vnímáním celkové kvality a srozumitelností řeči není zcela objasněn. Obecně platí, že řeč, která je vnímána jako „*dobrá*“ poskytuje vysokou srozumitelnost a naopak, viz. [1].

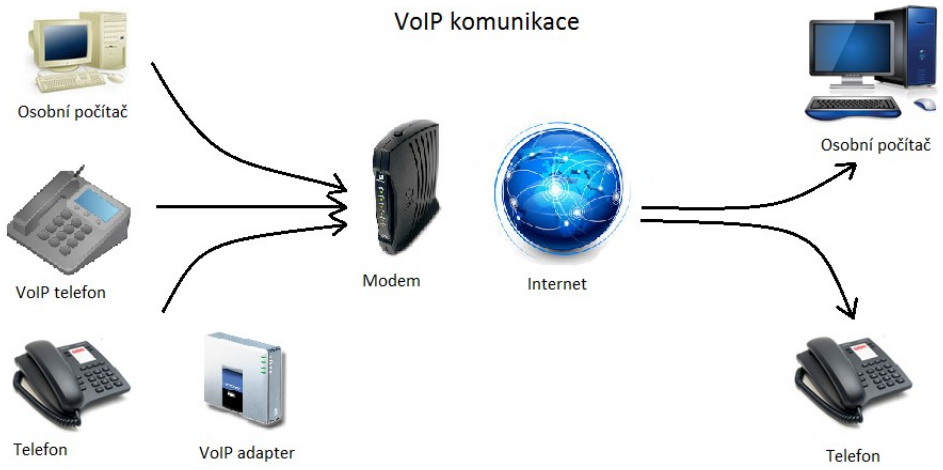


Obrázek 1: Kvalita řeči.

Další kritéria pro posuzování kvality řeči je lidské vnímání. To je jiné, pokud se jedná o muže či ženu, je-li řeč předem připravená a jen přečtená anebo když se jedná o řeč spontánní, kdy si ji řečník musí přímo na místě sám vymýšlet. Prostředí má vliv na řeč z důvodů, že se kolem nás vyskytuje neustále hluk. Měření hluku proto bylo prováděno ve venkovních prostorech, jako jsou hluky aut, tramvají nebo křižovatek, aby poté mohly být metody SNR vypočteny z hluků, které se vyskytují kolem nás. Více informací naleznete v [1].

1.1 Kvalita řeči ve VoIP

Kvalita řeči je pojem, který se začal používat, když se u klasických telekomunikačních sítí začaly vytvářet technologie, které měly těmto klasickým sítím konkurovat. Tyto nové technologie nabízely například nejrůznější vlastnosti spojení mezi dvěma volajícími jako: nové služby pro volající, levnější volání, mobilita účastníků a dále i kvalita služeb. Pojem kvalita služeb úzce souvisí s termínem kvalita řeči, která dominuje především v sítích VoIP. Komunikace přes internetovou síť je v časovém měřítku historická záležitost, avšak v dnešní době dochází k prudkému rozvoji tohoto druhu komunikace a kvalita řeči je velmi klíčová záležitost.



Obrázek 2: VoIP komunikace.

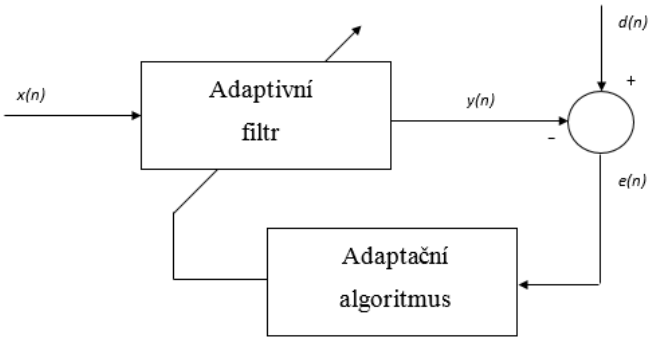
2 Zpracování řečového signálu

Řečový signál lze zpracovávat dvojím způsobem. Prvním je analýza dat, která jsou ve formě hotového záznamu, tedy offline přístup. Druhým způsobem je pořizování zvukové nahrávky, kdy výsledky analýzy jsou v krátkých časových intervalech zpracovávány, ukládány a aktualizovány, jedná se o online přístup.

Vstupní signál je nejprve filtrován a rozdělen na krátké časové úseky neboli segmenty, aby mohl být následně vyhodnocován detektorem řečové aktivity.

2.1 Adaptivní systémy

Adaptivní systém je tvořen adaptačním algoritmem, který určuje koeficienty adaptačního filtru. Základní adaptivní systém (Obr. 3) je tvořen čtyřmi signály: vstupní referenční signál $x(n)$, vstupní signál $d(n)$, výstupní chybový signál $e(n)$ a výstupní signál adaptivního filtru $y(n)$.



Obrázek 3: Základní adaptivní systém.

Adaptivní filtry se používají v aplikacích, které zahrnují kombinaci tří obecných problémů při zpracování signálu [2], [3]:

1. Odstranění hluku
2. Odhad průběhu signálu
3. Identifikace systému

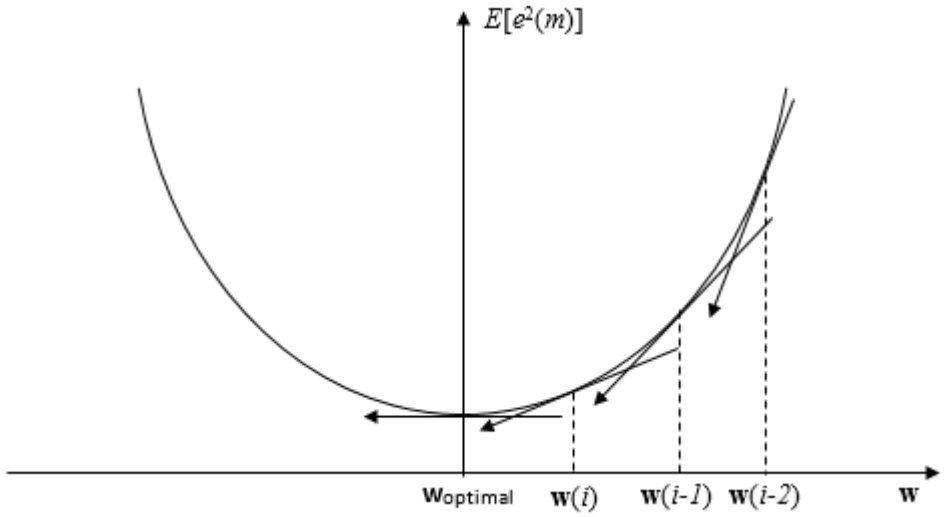
V této práci se budu zabývat aplikací adaptivních systémů pro odstranění hluku z nahrávek řeči.

2.1.1 Adaptační algoritmus LMS

Nejpoužívanější adaptivní algoritmus je LMS (metoda nejmenších čtverců). Tento algoritmus je také nazýván jako metoda největšího spádu. LMS bude výpočetně jednodušší metoda, když střední kvadratickou odchylku chybového signálu nahradíme samotným chybovým signálem. Adaptační algoritmus je definován jako [2], [3]:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu\left(\frac{\partial E[e^2(n)]}{\partial \mathbf{w}(n)}\right) = \mathbf{w}(n) + \mu\left(\frac{\partial e^2(n)}{\partial \mathbf{w}(n)}\right) \qquad (2.1)$$

Kde $\mathbf{w}(n)$ je vektor koeficientů adaptivního filtru, μ je označován jako velikost kroku. Jedná se o malou kladnou konstantu, která ovlivňuje stabilitu filtru, rychlost konvergence. Chybový signál $e(n)$ je rozdíl mezi vstupním signálem $d(n)$ a výstupním signálem $y(n)$. Na obrázku (Obr. 4) je znázorněna střední kvadratická odchylka chybového signálu, viz. [2].



Obrázek 4: Vyhledávání nulového gradientu střední kvadratické odchylky.

Chybový signál můžeme vyjádřit jako:

$$e(n) = d(n) - \mathbf{w}^{\mathrm{T}}(n)\mathbf{x}(n) \tag{2.2}$$

Při dosazení rovnice (2.2) do rovnice (2.1) dostaneme po následné úpravě rovnici:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu[\mathbf{x}(n)e(n)] \tag{2.3}$$

Lze vidět, že výsledná rovnice pro adaptivní algoritmus LMS je velmi jednoduchá. Hlavní výhodou tohoto algoritmu je jeho jednoduchost, jak z hlediska paměti, tak i z hlediska výpočetní náročnosti, viz. [2].

2.1.2 Adaptační algoritmus RLS

Tento algoritmus využívá jako kriteriální funkci sumu váhovaných čtverců chybového signálu. Algoritmus RLS má poměrně velkou míru konvergence k optimálním koeficientům. To je užitečné v aplikacích sloužících pro zpracování řeči, kde je nutné reagovat na rychlé změny signálu v čase [3].

Chybový signál je vyjádřen stejným vztahem jako u algoritmu LMS tedy (2.2).

Koeficienty filtru jsou adaptovány podle rovnice:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mathbf{k}(n)e(n) \tag{2.4}$$

Kde **k***(n)* je ziskový vektor, který se spočítá podle vztahu:

$$\mathbf{k}(n) = \frac{\lambda^{-1}\boldsymbol{\Phi}_{xx}(n-1)\mathbf{x}(n)}{1+\lambda^{-1}x^{T}(n)\boldsymbol{\Phi}_{xx}(n-1)\mathbf{x}(n)}, \tag{2.5}$$

kde $\boldsymbol{\Phi}_{xx}(n)$ je inverzní autokorelační matice referenčního signálu *x*(*n*). Adaptace matice se provádí pomocí rovnice:

$$\boldsymbol{\Phi}_{xx}(n) = \lambda^{-1}\boldsymbol{\Phi}_{xx}(n-1) - \lambda^{-1}\mathbf{k}(n)x^{T}(n)\boldsymbol{\Phi}_{xx}(n-1), \tag{2.6}$$

kde λ je faktor zapomínání, který se volí v rozmezí <0 1>. V praxi se tato hodnota volí mezi 0,98 až 1, viz. [2].

2.2 Segmentace, normalizace

Vzhledem k tomu, že řečový signál je brán jako stochastický, musí být před dalším zpracováním rozdělen na jednotlivé segmenty neboli rámce, které budou mít stejnou délku. Parametry těchto segmentů se mění velmi pomalu. Tyto segmenty jsou reprezentovány většinou časovým úsekem od 10 až 30 milisekund. V této bakalářské práci bude možnost volby v rozmezí od 20 do 32 milisekund. Segmentace je prováděna formou prostého dělení nahrávky na rámce délky *l* vzorků z celkového počtu vzorků *N, v*iz. [4], [13].

Před zpracováním signálu je možnost provést pár užitečných operací, jimiž jsou ustředění a normalizace signálu. Ustředění řečového signálu lze provést odečtením střední hodnoty signálu od každého vzorku, a tím se tedy odebere stejnosměrná složka signálu. Normalizací signálu dosáhneme toho výsledku, že výsledný signál bude mít rozsah amplitudy od <-1 1>. Nejprve se v signálu najde nejvyšší absolutní hodnota amplitudy signálu a touto hodnotou se potom provede podělení každého vzorku [5].

2.3 Krátkodobá energie

Krátkodobou energii signálu lze definovat vztahem:

$$E_n = \sum_{n=-\infty}^{\infty} [s(n)w(k-n)]^2, \tag{2.7}$$

kde *s*(*n*) je vzorek signálu v čase *n* a *w*(*k*) je příslušný typ okénka. Jedním z nedostatků této charakteristiky je značná citlivost na velké změny úrovně signálu. Z těchto důvodů se velmi často využívá krátkodobá intenzita, který zmíněný nedostatek nemá:

$$M_n = \sum_{n=-\infty}^{\infty} |s(n)|w(k-n). \tag{2.8}$$

Hodnoty krátkodobé intenzity mohou být využity například při automatickém oddělování segmentů ticha od segmentů řeči, viz. [4].

2.4 Autokorelační funkce

Krátkodobá autokorelační funkce je definována jako:

$$R_n(m) = \sum_{n=-\infty}^{\infty} s(n)w(k-n)s(n+m)w(k-n-m), \tag{2.9}$$

kde $s(n)$ je hodnota k-tého vzorku korelovaného signálu, m je hodnota posunu signálu o m vzorků a $w(k)$ je příslušný typ okénka.

Jestliže je zpracovávaný signál periodický s periodou P, pak autokorelační funkce nabývá maximálních hodnot právě pro m=0, P, 2P, …. Z uvedeného důvodu je charakteristika velmi vhodná k určování periody základního hlasivkového tónu, viz. [4].

2.5 Lineární prediktivní analýza

Lineární prediktivní analýza neboli lineární prediktivní kódování (LPC) je jednou z nejefektivnějších metod analýzy akustického signálu. Je to metoda, která se snaží odhadnout přímo z řečového signálu parametry modelu vytváření řeči. Atraktivnost této metody spočívá v její schopnosti zabezpečit velmi přesné odhady uvedených parametrů při přijatelné výpočetní zátěži. Princip metody LPC je založen na předpokladu, že n-tý vzorek signálu $s(n)$ lze popsat kombinací Q předchozích vzorků a buzení $u(n)$ [4]:

$$s(n) = -\sum_{i=1}^{Q} a_i s(n-i) + Gu(n), \tag{2.10}$$

kde G je koeficient zesílení a Q je řád modelu, prediktoru. Přenosovou funkci modelu $H(z)$ lze pak zapsat ve tvaru:

$$H(z) = \frac{G}{1 + \sum_{i=1}^{Q} a_i z^{-1}}. \tag{2.11}$$

Sledovanými parametry jsou zde koeficienty a_i číslicového filtru a koeficient zesílení G. Při splněném předpokladu stacionarity signálu na sledovaném časovém intervalu, lze pro výpočet koeficientů použít metodu nejmenších čtverců. Při neznámém členu $Gu(n)$ v rovnici (2.10), je z rovnice vypuštěn a vzniká chyba predikce $e(n)$ mezi skutečnou hodnotou $s(n)$ a předpovězenou $\hat{s}(n)$. Jedná se o krátkodobou energii chyby signálu danou vztahem [4]:

$$E = \sum_n e^2(n) = \sum_n [s(n) - \hat{s}(n)]^2 = \sum_n \left[s(n) + \sum_{i=1}^{Q} a_i s(n-i) \right]^2 \tag{2.12}$$

Úpravou rovnice (2.12) podle [4] lze dospět k maticovému zápisu:

$$\begin{bmatrix} R_n(0); & R_n(1); & R_n(2); & \cdots; & R_n(Q-1); \\ R_n(1); & R_n(0); & R_n(1); & \cdots; & R_n(Q-2); \\ \vdots & & & & \\ R_n(Q-1); & R_n(Q-2); & R_n(Q-3); & \cdots; & R_n(0); \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_Q \end{bmatrix} = \begin{bmatrix} -R_n(1) \\ -R_n(2) \\ \vdots \\ -R_n(Q) \end{bmatrix} \tag{2.13}$$

Kde $R_n(0)$ až $R_n(Q)$ jsou autokorelační koeficienty *n*-tého segmentu řečové aktivity. Matice na levé straně je v Töplitzově tvaru. Minimální střední kvadrát chyby predikce lze vyjádřit [4]:

$$E_n = R_n(0) + \sum_{i=1}^{Q} a_i R_n(i) \tag{2.14}$$

2.6 Levinson-Durbin

Z matice (2.13), která je symetrická, můžeme dále počítat soustavu rovnic, kterou navrhnul Levinson a modifikoval Durbin, jejíž výsledkem bude matice v takovém tvaru, že všechny prvky pod diagonálou budou nulové.

V Durbinově metodě je řešení pro soustavu rovnic vyjádřeno rekurzivně pro *i*=1, 2, …*Q* jako:

$$\begin{aligned} & E_n^{(0)} = R_n(0), \\ & k_i = -\left[R_n(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R_n(i-j) \right] / E_n^{(i-j)} \\ & a_i^{(i)} = k_i \\ & a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}, \qquad 1 \le j \le i-1 \\ & E_n^{(i)} = (1 - k_i^2) E_n^{(i-1)} \end{aligned} \tag{2.15}$$

Kde $a_j^{(i)}$je *j*-tý parametr prediktoru řádu *i* a *Q* představuje řád prediktoru a obecně se volí podle vzorce:

$$Q = \frac{f_{vz}}{1000} + 4 \tag{2.16}$$

Příklad použitý v této práci bude při vzorkovací frekvenci 51200 Hz hodnota prediktoru řádu *Q* rovna 55 viz [4], [11] a [12].

2.6.1 PARCOR koeficienty

V praxi se velmi často místo koeficientů a_i využívá mezivýsledků Durbinova algoritmu, tj. tzv. koeficientů odrazu $k_i (1 \le i \le Q)$. V literatuře zabývající se statistikou jsou tyto koeficienty známy pod označením PARCOR (částečné korelační koeficienty). Pro tyto koeficienty platí podmínka

$k_i \in \langle -1,1 \rangle$. Hodnoty energie chyby predikce $E_n^{(i)}$ získáváme jako mezivýsledky při výpočtu koeficientů a_i, viz. [4], [11].

$$\begin{matrix} a_1^{(1)} & a_1^{(1)} & a_1^{(1)} & \cdots & a_1^{(Q)} \\ 0 & a_2^{(1)} & a_2^{(1)} & \cdots & a_2^{(Q)} \\ 0 & 0 & a_3^{(1)} & \cdots & a_3^{(Q)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_Q^{(Q)} \end{matrix} \tag{2.17}$$

Jako PARCOR koeficienty jsou vybrány hodnoty diagonály, tedy $a_1^{(1)}$ až $a_Q^{(Q)}$.

3 Hodnocení kvality řeči

Obecně platí, že kvalita řeči se dá hodnotit pomocí subjektivních a objektivních metod. Subjektivní metody jsou založeny na poslechovým nebo konverzačních testech, kdy posluchači hodnotí kvalitu promluv. Posluchači hodnotí kvalitu řeči podle předem stanovené stupnice subjektivně. Výsledky hodnocení jednoho posluchače budou ve většině případů rozdílné. Tato odchylka může být snížena průměrem výsledků více posluchačů. Objektivní metody využívají matematických algoritmů, které se snaží předpovídat subjektivní hodnocení zkoumáním digitalizovaného řečového signálu. Typicky se vypočítají jako určitá vzdálenost mezi referenční řečí a zkresleným projevem, viz. [1]

3.1 Subjektivní hodnocení kvality řečových signálů

Subjektivní metody hodnocení kvality jsou určeny pro obecné použití, jelikož jsou nezávislé na typu degradace signálu, ať se jedná o hluk, ozvěny nebo chybovost přenosu. Metody jsou detailně popsány v doporučeních ITU-T řady P [6], [7], [8].

K subjektivnímu hodnocení kvality telekomunikačních zařízení a systémů lze použít dva typy metod – poslechové a konverzační.

3.1.1 Konverzační metody

Konverzační testy jsou určeny k laboratorní simulaci situací, se kterými se setkávají uživatelé telekomunikačních koncových zařízení. Při použití konverzačních testů je nutné zvolit vhodné podmínky a testovací subjekty, stejně důležité je i korektní vedení testu.

- Testovací místnost

Dva subjekty sedí v oddělených zvukotěsných místnostech blízko místa, ze kterého je experiment kontrolován. Místnost musí mít více než 20 metrů krychlových s dobou ozvěny menší než půl sekundy (obvykle v rozmezí 200–300 milisekund). Hluk v místnostech se musí udržovat na co nejnižší možné míře.

Fyzická konstrukce místnosti by měla být taková, aby dostatečně utlumila zvuk z vnějšího prostředí.

- Stupnice názorů na konverzaci

Různé pětibodové stupnice kategorií hodnocení mohou být užity k různým účelům. Rozsah a formulace názorových stupnic, předložených subjektům při experimentu, hrají důležitou roli a měly by se držet standardu prověřeného již provedenými experimenty. Následující názorové rozvržení stupnice je nejpoužívanější pro ITU-T aplikace a mely by být použity stejné doslovné ekvivalenty, více v [6], [9] a [10].

Tabulka 1: Škála MOS pro hodnocení kvality [9]

	Škála MOS
5	Excellent – vynikající kvalita
4	Good – dobrá kvalita
3	Fair – přijatelná kvalita
2	Poor – špatná kvalita
1	Bad – velmi špatná kvalita

3.1.2 Poslechové metody

Výsledky poslechových testů nejsou stejně realistické, jako testy konverzační. Je to z toho důvodu, že kritéria poslechových testů jsou méně přísná. Umělost tohoto testu znamená, že je nutné sledovat, kontrolovat a specifikovat mnoho parametrů, které jsou při konverzačních testech automaticky v rovnováze. Proces testování je však jednodušší než u testů konverzačních.

- Testovací místnost

Testovací subjekty by měly být umístěny do zvukotěsné místnosti, která bude mít velikost v rozmezí od 20 do 120 metrů krychlových s dobou ozvěny menší půl sekundy (obvykle v rozmezí 200–300 milisekund). Tato místnost by měla být také bez zvuková. V menších místnostech může dojít ke zkreslení zvuku vlivem odrazu. Hluk místnosti musí být udržován na co nejnižší možné míře.

- Nahrávací systém

  Záznamový systém musí mít vysokou kvalitu a může jím být:

  - Konvenční dvoukanálový rekordér,
  - Dvoukanálový digitální zvukový procesor s vysoce kvalitním videorekordéru nebo Digital Audio Tape (DAT),
  - Počítačem řízený digitální úložný systém.

Systém řízený počítačem je nejvhodnější a nejvšestrannější, ale z praktických důvodů často diktuje volbu jednoho z ostatních systémů.

- Řečový materiál

Řečový materiál by se měl skládat z jednoduchých, smysluplných, krátkých vět náhodně vybraných pro snadné pochopení. Tyto věty by měly být v náhodném pořadí tak, že jedna věta nenavazuje na druhou. Dlouhým větám je třeba se vyvarovat. Každá věta by měla být v úseku 2-3 sekund. Příklady pro řečový materiál:

„Budeš muset být velmi tichý.“

„Nebylo tu nic k vidění.“

„Potřebuješ nějaké peníze?“

Experimentátor musí rozhodnout, kolik vět je nutné v každé skupině představit. Doporučuje se minimálně dvě věty a maximálně pět vět.

- Nahrávací procedura

Pro nahrávání řeči se používá lineární mikrofon a nízko-šumový zesilovač s plochou frekvenční odezvou. Mikrofon je umístěn ve vzdálenosti 140 až 200 mm od rtů mluvčího. Lze použít současně dva samostatné záznamové systémy: jeden pro záznam širokopásmové řeči a druhý pro záznam telefonní řeči. Tento systém se dvěma záznamy zajišťuje, že stejná řeč je zaznamenávána ve dvou formách (širokopásmová řeč a telefonní řeč). Za normálních okolností se vyžaduje pouze jeden záznamový systém, ale existují případy, kdy je nutné použít obě metody záznamu, a to je výhodné v každém případě, jelikož je možné provést srovnávací měření na dvou verzích. Během záznamového procesu se udržuje hodnota aktivní řečové úrovně mezi 20 a 30 dB pod úrovní přetížení nahrávacího systému, více v [6] a [9].

3.1.3 DTW (Dynamic time warping)

Metoda borcení časové osy (Dynamic Time Warping) DTW, využívá principu dynamického programování. Tato metoda sice patří mezi starší, avšak pro svou jednoduchost je stále využívána. Slouží k rozpoznávání izolovaných slov (signálů) nebo krátkých úseků. Tento algoritmus se používá zejména pro rozpoznávání řeči a izolovaných, nebo klíčových slov. Jako vzor pro rozpoznávání se používá ve většině případů referenční nahrávka. Postupně, pomocí algoritmu, lze porovnat řečníkem vyslovené slovo (úsek slov) s danou referenční nahrávkou a spočítat vzdálenost cesty DTW. Hledá se ten referenční signál, který měl s daným testovaným signálem nejmenší vzdálenost. Práce se bude zabývat využitím této metody při klasifikaci nahrávek řeči se stejnými nahrávkami, které byly doplněny o různý šum a následně odfiltrovány pomocí adaptivní filtrace a tato odfiltrovaná řeč je použita jako testovaná.

Každý signál je vyjádřen posloupností vektorů. Jako testovaný signál budeme považovat signál **A**:

$$\mathbf{A} = \{\mathbf{a}(1), \mathbf{a}(2), \mathbf{a}(3), \dots, \mathbf{a}(n), \dots, \mathbf{a}(I)\}, \tag{3.1}$$

a jako referenční signál pod písmenem **B**:

$$\mathbf{B} = \{\mathbf{b}(1), \mathbf{b}(2), \mathbf{b}(3), \dots, \mathbf{b}(m), \dots, \mathbf{b}(J)\}. \tag{3.2}$$

Znamená to tedy, že **a**(n) značí n-tý vektor testovaného signálu **A** a **b**(m) je naopak m-tý vektor referenčního signálu **B**. DTW hledá v rovině (n,m) optimální cestu

$$m = \mathbf{\Psi}(n). \tag{3.3}$$

Tato cesta minimalizuje funkci D, což je vzdálenost mezi jednotlivými vektory signálů **A** a **B**:

$$D(\mathbf{A}, \mathbf{B}) = \sum_{n=1}^{I} \hat{d}\left[\mathbf{a}(n), \mathbf{b}\left(\mathbf{\Psi}(n)\right)\right], \tag{3.4}$$

kde $\hat{d}[\mathbf{a}(n), \mathbf{b}(\boldsymbol{\Psi}(n))]$ je lokální vzdálenost mezi n-tým vektorem testovaného slova a *m*-tým vektorem referenčního slova. Je zavedena časová proměnná *k*. Časové proměnné *m* a *n* jsou takovými funkcemi *k*, že platí:

$$n = i(k),$$

$$m = j(k), \tag{3.5}$$

V případě, že jsou dva signály, u kterých známe počáteční a koncové body, můžeme vyjádřit omezení funkci DTW hraničními body,

$$\begin{aligned} i(1) &= 1, \qquad j(1) = 1, \\ i(K) &= I, \qquad j(K) = J. \end{aligned} \tag{3.6}$$

Jednotlivé prvky matice **G** jsou rekurzivně vypočteny dle vztahu:

$$g[i(k), j(k)] = \min_{\{i(k), j(k)\}} \{g[i(k-1), j(k-1)] + d[i(k), j(k)]\widehat{W}(k)\}, \tag{3.7}$$

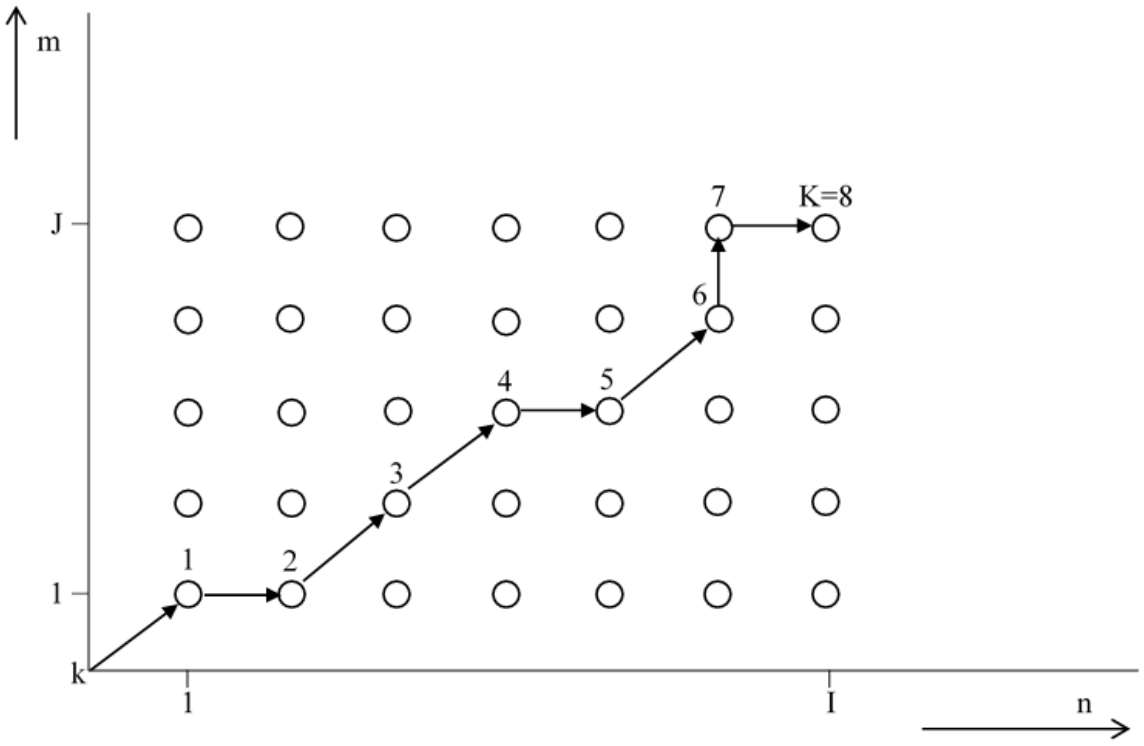
kde $k = 1, \ldots, K$ a $d[i(k), j(k)]$ je příslušná lokální vzdálenost z matice **D**. Je-li do vztahu (3.9) dosazena váhová funkce typu a), pak lze odvodit konkrétní vztah pro výpočet lokálního omezení typu I, který je dán vztahem:

$$g(n,m) = min \begin{cases} g(n, m-1) + d(n,m) \\ g(n-1, m-1) + 2d(n,m) \\ g(n-1, m) + d(n,m) \end{cases} \tag{3.8}$$

Konečnou normalizovanou vzdálenost mezi obrazy **A** a **B** lze vyčíslit dle vztahu:

$$D(\mathbf{A}, \mathbf{B}) = [N(\widehat{W})]^{-1} g[i(K), j(K)] = \left[N\left(\widehat{W}\right)\right]^{-1} g[I, J] \tag{3.9}$$

Bude-li hodnota **D** rovná 0 znamená to, že obrazy **A** a **B** mají největší shodu, takže jsou si nejvíce podobné, v opačném případě pak bude mít hodnotu 1, [4], [5], [11].



Obrázek 5: Funkce DTW pro testovaný a referenční obraz v rovině (n,m).

3.2 Objektivní hodnocení kvality řečových signálů

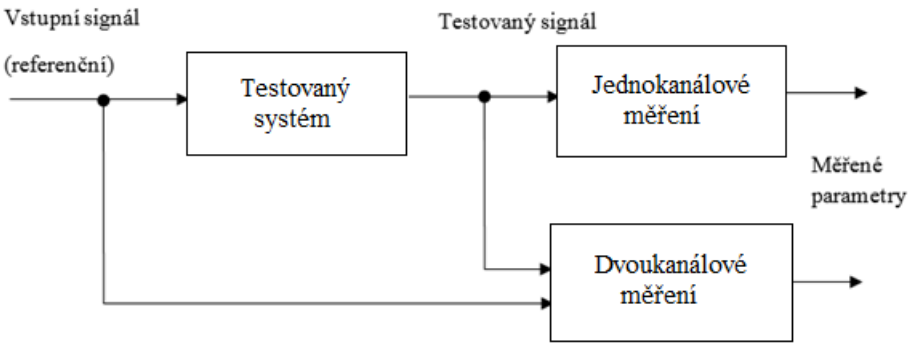
Subjektivní metody hodnocení kvality řeči patří mezi nejspolehlivější metody, jelikož samotný posluchač hodnotí kvality řeči a její srozumitelnost. Nicméně tyto metody jsou časově velice náročné a vyžadují, ve většině případů, testované a trénované posluchače.

Z těchto důvodů byly vytvořeny objektivní metody hodnocení kvality řeči, které jednodušším způsobem odhadnou subjektivní vlastnosti řeči.

Pro získání korelačních koeficientů pro subjektivní a objektivní metody musí být použita stejná databáze testovaných signálů. Korelační analýza je využita pro zjištění, jestli objektivní metody předpovídají kvalitě řeči tak, jako by ji hodnotili posluchači pomocí testů subjektivních, [1], [7], [10].

Objektivní metody lze rozdělit na dvě skupiny:

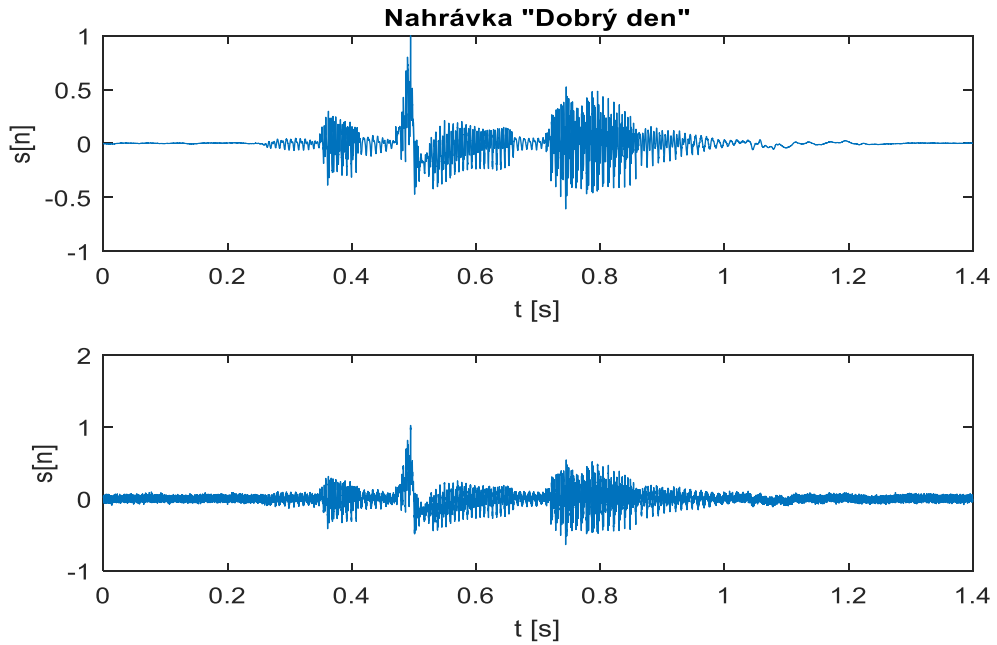
- Jednokanálové měření
  U těchto měření máme k dispozici pouze degradovaný nebo zvýrazněný řečový signál. Řečový signál rozdělíme na segmenty pevné délky (většinou 10–30 milisekund), poté provedeme výpočty pro každý rámec a výsledek se vytvoří z hodnot vypočtených v jednotlivých rámcích.
- Dvoukanálové měření
  Porovnávají originální signál a degradovaný, který prošel testovaným systémem.



Obrázek 6: Zobrazení jednokanálové a dvoukanálové metody.

3.2.1 Odstup signálu od šumu (SNR)

Odstup signálu od šumu (Signal-to-Noise Ratio, SNR) je velmi často používaná veličina v elektronice. Je důležitou kriteriální veličinou na bázi výkonových charakteristik signálu, kde se používá ke kvalifikaci úrovně aditivního šumu (rušivého signálu) v analyzovaném užitečném signálu. Z hlediska nejčastější aplikace měření odstupu dvou akustických signálů vychází definice SNR z vlastností vnímání intenzity zvuku lidským sluchovým ústrojím, které není lineární, ale logaritmické. Je definován jako poměr výkonu užitečného signálu, čímž se rozumí řečový signál, a výkonu neužitečného signálu, tedy hluku. Čím větší odstup mezi těmito signály bude, tím bude výsledný signál kvalitnější, bude obsahovat méně šumu. Jednotkou poměru signálu od šumu je decibel (dB). Výhodnou této metody je, že je výpočetně jednoduchá.



Obrázek 7: První časový průběh je nezkreslené slovní spojení a druhý průběh téhož signálu se superponovaným Gaussovým šumem

Metoda SNR není příliš vhodná pro určení kvality řečového signálu. Výsledky se příliš neshodují s výsledky dosaženými subjektivním měřením. Hlavním důvodem je, že řečový signál je nestacionární.

Chybový signál je definován následující rovnicí:

$$x(n) = s(n) + noise(n). \tag{3.10}$$

Kde $x(\mathrm{n})$ vzorek směsi řečového signálu, $s(n)$ řečový signál bez hluku, $noise(n)$ signál hluku. Všechny signály mají stejnou délku *l*.
Základní definice odstupu signálu od šumu pro deterministické signálu je dána vztahem:

$$SNR = 10 \log \frac{P_s}{P_n}, \tag{3.11}$$

kde P_s je výkon užitečného signálu a P_n je výkon hluku.

Definice odstupu signálu pro náhodné signály je dána vztahem:

$$SNR = 10 \log \frac{\sigma_s^2}{\sigma_n^2}. \tag{3.12}$$

Řečový signál i šumové pozadí jsou signály s nulovou střední hodnotou, a proto jsou výkony signálů dány jejich rozptyly σ_s^2 a σ_n^2 .

Pro deterministické signály je SNR definováno na bázi výkonu signálů, viz. (3.11). Pro náhodné signály je SNR definováno na bázi rozptylu signálu.

Výše popisované definice (3.11) a (3.12) jsou často zaměňovány zejména v případech, kdy pracujeme se signály s nulovou střední hodnotou.

Pro zapamatování několik konkrétních faktů zřejmé z definice SNR podle (3.11):

- Kladné SNR znamená, že užitečný signál je silnější než šum, záporné SNR naopak, tj. hluk je silnější než užitečný signál.
- Hodnota 10 dB znamená řádový rozdíl v SNR, tj. signál je z hlediska výkonu 10krát silnější než hluk, a tedy to bude platit i naopak pro -10 dB, že signál je 10krát slabší než hluk.
- Hodnota 6 dB připadne situaci, kdy výkon signálu je dvojnásobný než výkon šumu.
- Hodnota 3 dB, znamená, že je poměr výkonu signálu a výkonu roven $\sqrt{2}$.

Více informací naleznete v [10], [14] a [15].

3.2.2 Globální SNR (GSNR)

Globální SNR (GSNR) dostaneme aplikací (3.11) na řečový signál, počítáme-li výkony řeči a šumu přes celý signál:

$$GSNR = 10\log\frac{\sigma_s^2}{\sigma_n^2} = 10\log\frac{\sum_{n=0}^{l-1} s^2(n)}{\sum_{n=0}^{l-1} noise^2(n)}, \tag{3.13}$$

kde l je délka řečového signálu ve vzorcích, $s(n)$ je vzorek řeči a $noise(n)$ je vzorek šumu. Pro globální SNR je charakteristické, že s analyzovaným signálem pracuje jako s celkem. Toto kritérium je však zatíženo chybou, protože do výpočtu výkonu řeči jsou zahrnuty i části signálu bez řečové aktivity, které snižují celkový výkon řečového signálu. V řečových pauzách je teoreticky nulový výkon signálu.

Ve skutečnosti je i zde zbytkový výkon generovaný hlasovým ústrojím člověka. Korektní výpočet *SNR*‘ pro řečový signál je definován takto:

$$SNR' = 10\log\frac{\sum_{n=0}^{l-1} s^2(n) * vad(n)}{\sum_{n=0}^{l-1} noise^2(n) * vad(n)}, \tag{3.14}$$

kde $vad(n)$ je informace o řečové aktivitě pro daný vzorek signálu (0 – pauza, 1 – řeč). Tento vzorec již respektuje pauzy v řečovém signálu, [10], [14] a [15].

3.2.3 Lokální SNR

Protože je řečový signál nestacionární, není SNR v závislosti na čase konstantní. Často nás může zajímat právě vývoj SNR v závislosti na čase. Rozdělíme-li řečový signál na úseky o délce maximálně 32 milisekund, můžeme řeč v těchto úsecích považovat za stacionární. Mluvíme tak o kvazistacionaritě řeči. Délka segmentu řeči je při vzorkovacím kmitočtu signálu *fs* = 51200 *Hz* zvolena na 1600 vzorků, aby odpovídala délce 32 milisekund. Jednotlivé segmenty na sebe mohou navazovat nebo se mohou překrývat. Překryv je typicky poloviční. Většinou pak vystačíme s vyčíslením SNR pro tyto segmenty.

Výsledkem je tak *lokální* SNR (*LSNR*), definované pro i-tý segment jako:

$$SNR_i = 10\log\frac{\sigma_s^2}{\sigma_n^2} = 10\log\frac{\sum_{n=0}^{N-1} {s_i}^2(n)}{\sum_{n=0}^{N-1} {n_i}^2(n)}, \tag{3.15}$$

kde $s_i(n) = s(m\,i + n), n_i(n) = n(m\,i + n)$, N je délka segmentu a m krok segmentace. V řečových pauzách je teoreticky $SNR_i = -\infty$. V případě reálného signálu je v řečových pauzách přítomen zbytkový signál. Hodnota SNR_i se pak v těchto segmentech obvykle nahrazuje zápornou hodnotou typicky -40 dB, [10], [14] a [15].

3.2.4 Segmentální SNR (SSNR)

Segmentální SNR (SSNR) má oproti ostatním metodám SNR tu výhodu, že výpočet se neprovádí přes celý zašuměný signál, ale berou se jen ty úseky, ve kterých se vyskytuje řečová aktivita.

Proto se tedy při výpočtu SSNR z řečového signálu setkáme s požadavkem na detekci těchto úseků. Tato detekce se dá provádět pomocí detektoru řečové aktivity (VAD – Voice Activity Detector). Detektor řečové aktivity je vysvětlen níže.

SSNR pak počítáme podle vztahu:

$$SSNR = \frac{1}{K_{VAD}} \sum_{i=0}^{L-1} \left(10 \log \frac{\sum_{n=0}^{N-1} {s_i}^2(n)}{\sum_{n=0}^{N-1} noise^2(n)} * VAD_i \right), \tag{3.16}$$

kde L je celkový počet analyzovaných segmentů, N je délka segmentu a K_{VAD} je počet segmentů s řečovou aktivitou. Hodnota VAD_i nese informaci o řečové aktivitě v i tém segmentu (1- řeč, 0 - pauza). Bližší popis jednotlivých SNR definicí naleznete v [10], [14] a [15].

3.2.5 Detektor řečové aktivity

Princip stanovení řečového aktivity je u většiny detektorů obdobný a lze jej shrnout do několika kroků:

1. Vstupní signál se rozdělí na stejné časové segmenty neboli rámce.
2. Stanovení potřebné charakteristiky signálu (dle zvoleného detektoru, např. intenzita, energie, kepstrum …).
3. Vypočítá se prahová hodnota charakteristiky. Prahová hodnota může být v průběhu detekce přepočítávána a aktualizována, nebo též stanovena pevně.
4. Vypočtená charakteristika se v každém rámci porovná s prahovou hodnotou.
5. Je-li daná charakteristika signálu v rámci (segmentu) větší než prahová, je detekována řeč, bude-li hodnota menší, než prahová je detekována pauza.

Inicializační fáze: inicializace probíhá na začátku signálu. V tomto úseku nesmí být přítomna řeč. Nastaví se hodnota prahu. Z několika počátečních rámců se vypočte střední hodnota M_n, která je brána jako hodnota prahová.

Fáze detekce: určí se hodnota vybraného rámce signálu. Je-li hodnota větší než prahová hodnota, je detekována řeč, jinak je detekována pauza a dojde k aktualizaci prahové hodnoty.

3.2.5.1 Typy detektorů:

Mnoho systémů, které odstraňují nežádoucí rušivý signál z řeči, vyžadují přesné stanovení úseku řeči a mezery v daném signálu. K tomuto účelu slouží různé detektory řeč/pauza.

Ideální detektor

Ideální detekce se realizuje tak, že ručně označíme úseky signálu s řečovou aktivitou. Pro svoji složitost se však tato detekce nedá použít při pořízení většího množství dat a taky jí nelze detekovat signál v reálném čase.

Energetický detektor

Tento detektor výpočtem zjišťuje energii signálu pro každý rámec. Využívá se funkce krátkodobé energie, tu lze definovat vztahem:

$$E_n = \sum_{n=-\infty}^{\infty} s(n)^2 \,, \tag{3.17}$$

kde $s(n)$ je vzorek signálu v čase n.

Při měření krátkodobé energie je doporučená délka rámců 20 až 32 milisekund. Funkce je citlivá na změny úrovně signálu, proto se často využívá krátkodobá intenzita, která tento nedostatek nemá:

$$M_n = \sum_{n=-\infty}^{\infty} |s(n)|. \tag{3.18}$$

V bakalářské práci je použit právě detektor řeči založený na výpočtech prahových hodnot intenzit.

Prahová hodnota se vypočítá jako suma počtu zvolených segmentů podělených počtem zvolených segmentů:

$$M_d = \frac{\sum_{x=0}^{X-1} M_x}{X}, \tag{3.19}$$

kde M_x je intenzita prvních x zvolených segmentů. Dále se vypočtený segment intenzity porovná s prahovou hodnotou, která je definována jako:

$$M_p = 1{,}5 M_d. \tag{3.20}$$

M_p dále považuji jako prahovou hodnotu, která se dá měnit pomocí počtu segmentů intenzit. Při detekci pauzy se provede aktualizace prahové hodnoty. Jedná se o vyhlazení intenzity jednotlivých segmentů, kde *M* je intenzita signálu v segmentu a parametr *p* je volí v rozmezí <0 1>, nejčastěji bývá volen od 0,4 až 0,6. Aktualizace je provedena podle:

$$M_d^{aktuální} = (1-p) M_d^{předchozí} + pM, \tag{3.21}$$

kde M_d je úroveň hluku pozadí.

Aktualizace se provádí jen v řečových pauzách. Je-li $M_n > M_p$ je detekována řeč, jinak je detekována řečová pauza, [12], [13] a [16].

4 Praktická část

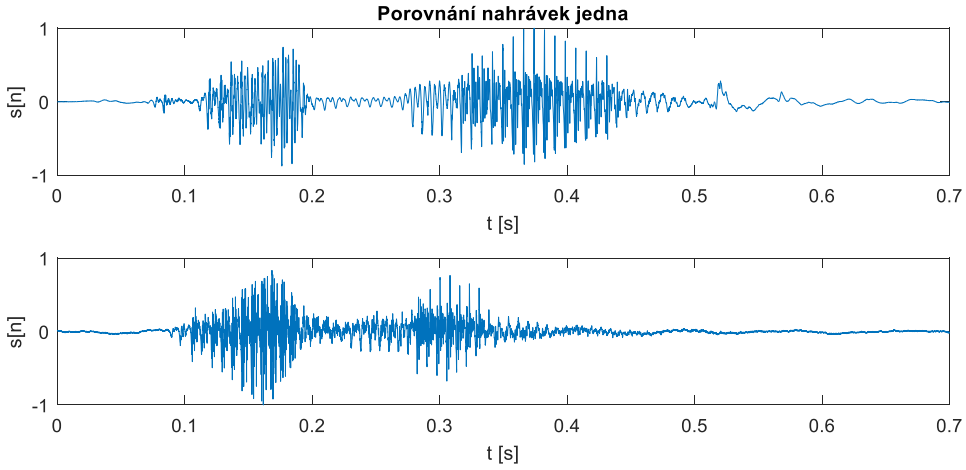
Výsledkem bakalářské práce jsou dva programy vytvořené v programovacím a vývojovém prostředí LabVIEW. První program slouží pro nahrávání vlastních nahrávek řeči a hluku z okolí. Nahrávky řeči byly pořízeny pomocí notebooku, měřicí karty od NI-9234 a mikrofonu G.R.A.S. 40PP CCP. Parametry jsou uvedeny v tabulce se srovnáním s řečí:

Tabulka 2: Parametry měřící karty a mikrofonu ve srovnaní s lidskou řečí.

	G.R.A.S. 40PP CCP	**NI-9234**	**Lidská řeč**
Frekvence (Hz)	10 Hz až 20 kHz	(Vzorkovací) 51,2 kHz	300 Hz až 4 kHz
Dynamický rozsah (dB)	135 dB	102 dB	96 dB

Podle Shannon-Kotělnikovova teorému, který nám říká, že při rekonstrukci signálu, musí být vzorkovací frekvence rovna minimálně dvojnásobku nejvyšší harmonické složky obsažené ve vzorkovaném signálu. Při nedodržení podmínky by docházelo k aliasingu. Podle tohoto teorému musí být vzorkovací frekvence pro řeč, stanovena na minimální hodnotu 8 kHz. Vzorkovací frekvence byla stanovena na hodnotu 51200 Hz podle měřící karty.

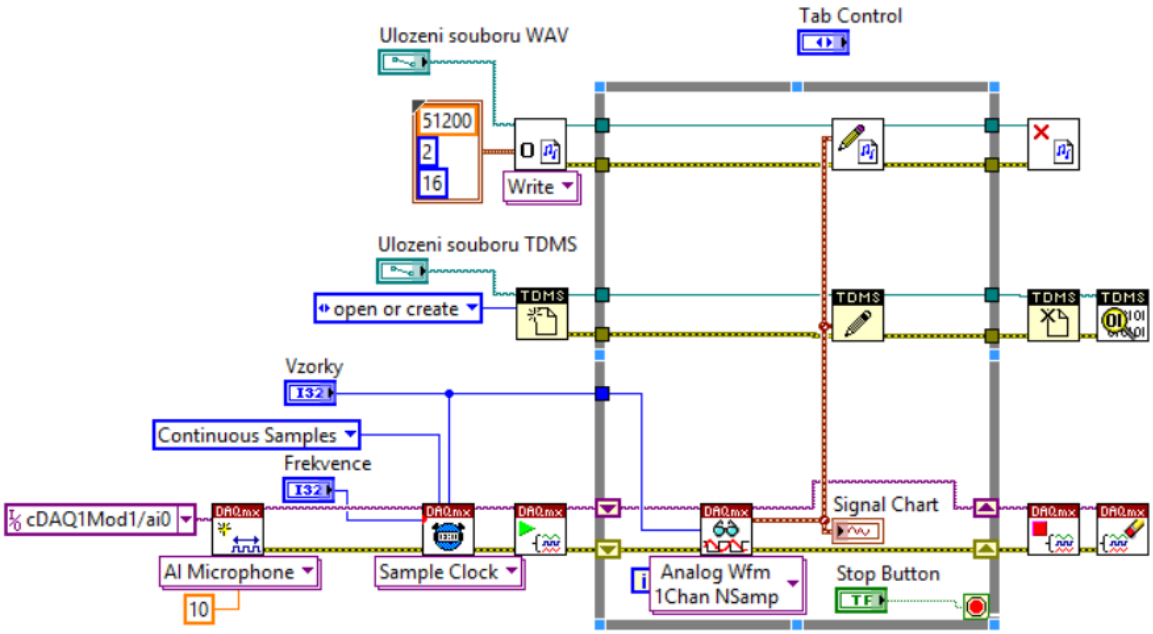
Vysoký dynamický rozsah měřícího mikrofonu 135 dB způsobuje, že při měření je třeba umístit mikrofon kousek od úst při nahrávání referenčních nahrávek, jinak bude docházet, z důvodu velkého dynamického rozsahu, k ovlivňování řeči okolním hlukem.



Obrázek 8: Porovnání dvou nahrávek "jedna".

V grafu (Obr. 8) je zobrazena citlivost mikrofonu při pořizování dvou stejných nahrávek slova „*jedna*“. První nahrávka byla pořízena při vzdálenosti 10 centimetrů úst od mikrofonu, tedy mikrofon byl kousek od úst. Na druhé nahrávce lze vidět, že když byl mikrofon vzdálen od úst 50 centimetrů, tak nahrávka již nemá čistý průběh, a tady se projevuje velký dynamický rozsah měřícího mikrofonu. Další nahrávky byly pořizovány při malé vzdálenosti úst od mikrofonu.

Program pro nahrávání a ukládání *Nahravani a ukladani dat* vyžaduje pro komunikaci s měřící kartou knihovnu *DAQmx Data Acquisition*. Ukládání je prováděno pomocí TDMS souboru, který byl použit z důvodů jeho kompatibility s různými druhy souborů, a dále pomocí WAV, který slouží pro přehrávání nahrávek. Zapojení je zobrazeno na obrázku 9:

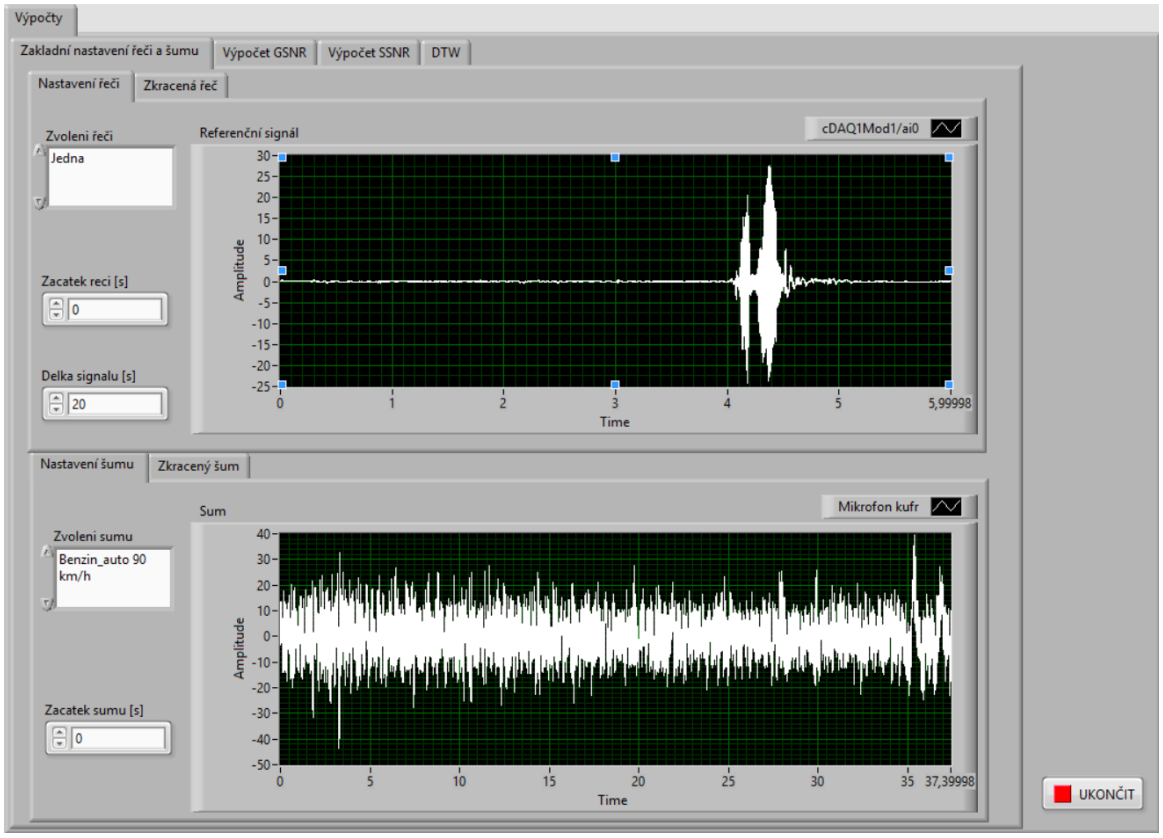


Obrázek 9: Zapojení pro nahrávání a ukládání dat.

V tomto programu se musí nejdříve vytvořit kanál, na který se připojí měřící mikrofon, dále se pomocí hodin nastaví vzorkovací frekvence podle karty a nepřetržité vykreslování vzorků.

Před samotným hodnocením signálu pomocí objektivních a subjektivních metod jsou nahrávky upravené tak, jak je popsáno v teoretické části. Tedy od nahrávek se odečte střední hodnota a následně se omezí amplituda v rozmezí od <-1 1>, tzn. že se provede ustředění a normalizace signálu.

Druhý program je *Vyhodnocovani kvality reci*. Program se ovládá pomocí grafického prostředí, kde není zapotřebí mít znalosti s programovacím prostředí LabVIEW. Výsledné uživatelské rozhraní je zobrazeno na obrázku 10:



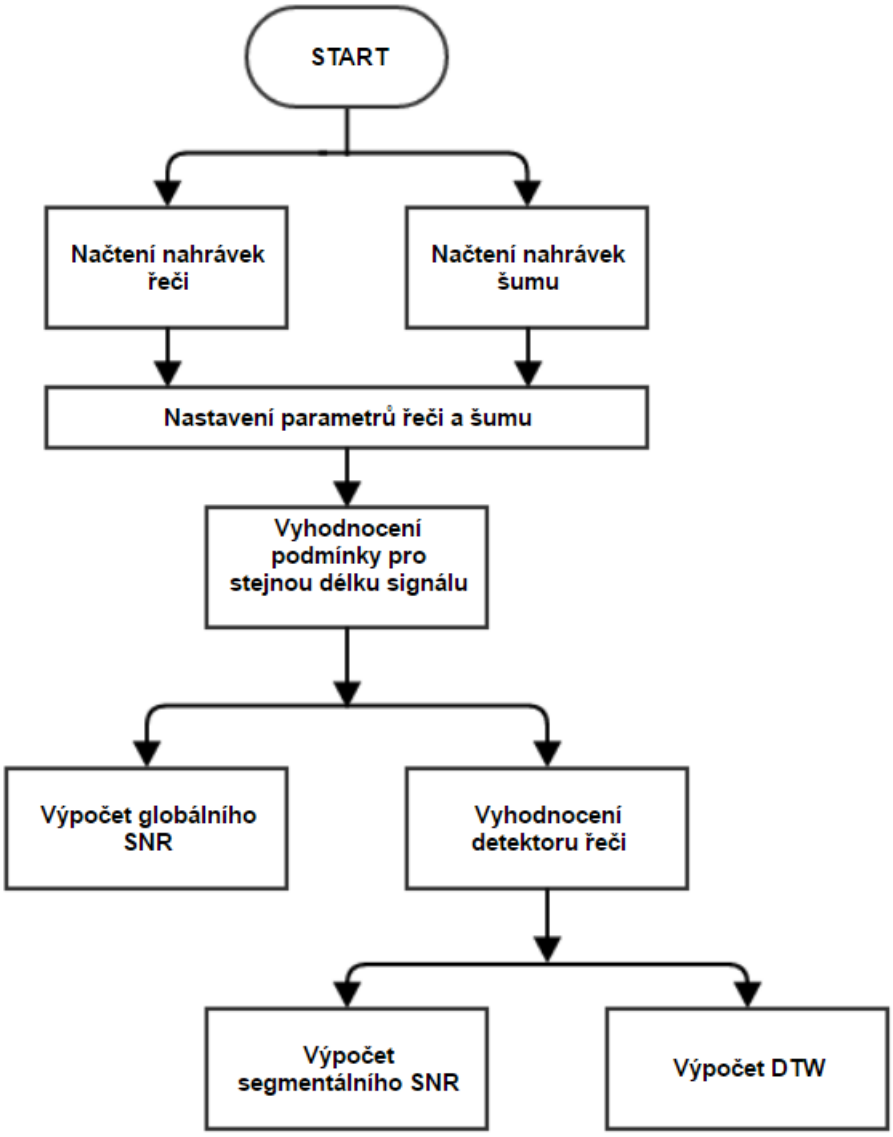
Obrázek 10: Grafické zobrazení výsledného programu.

Pro spuštění programu stačí stisknout tlačítko run. Na začátku si uživatel vybere z různých nahrávek řeči a hluku, které bude používat pro další zpracování a v grafu se vám ukáží jejich průběhy.

Po provedení výběru nahrávky, je zde možnost vybrat určitý úsek řeči/šumu pomocí zadání hodnot do *Zacatek reci, Zacatek sumu a Delka signalu*. Mezi tím dochází k neustálým výpočtům v programu:

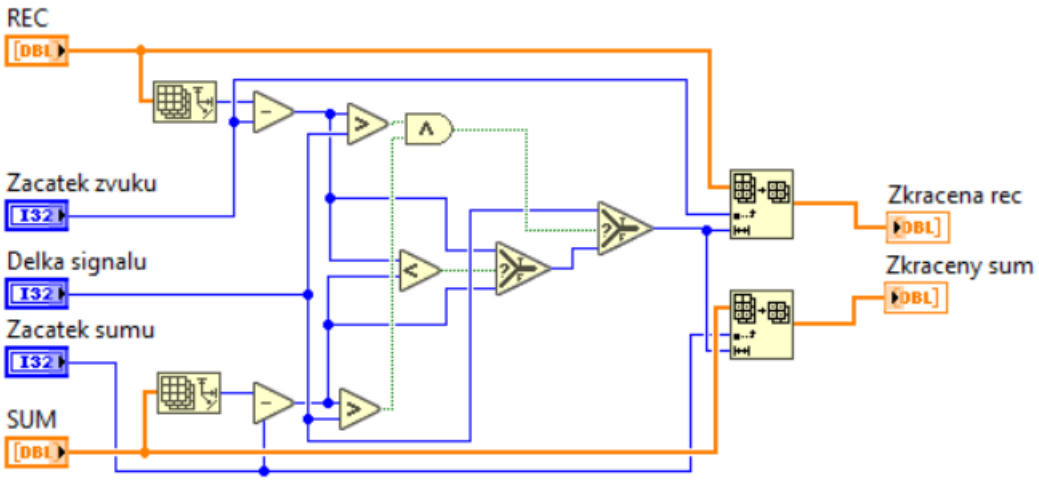
- GSNR Základní metoda výpočtu poměru signálu od šumu
- SSNR Segmentální SNR s využitím detektoru řečové aktivity
- DTW Výpočet vzdáleností mezi dvěma porovnávanými slovy

Na obrázku (Obr. 11) je zobrazen vývojový algoritmus, který probíhá neustále od spuštění programu do jeho vypnutí.



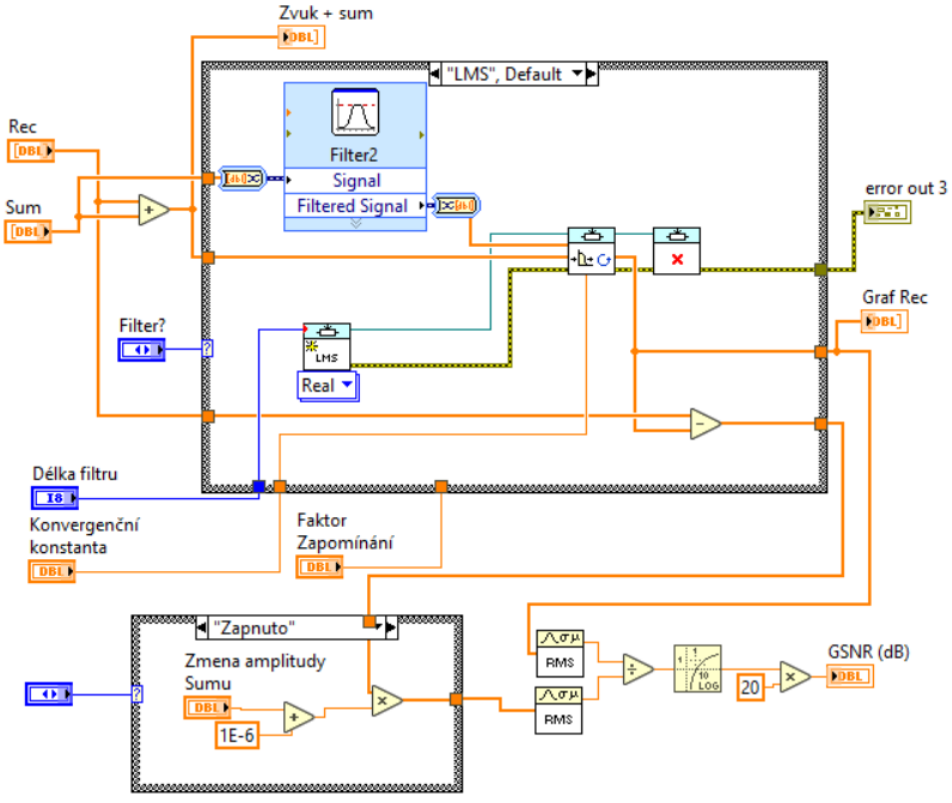
Obrázek 11: Vývojový algoritmus celého programu.

Z důvodů použití adaptivní filtrace pro odfiltrován zašuměného signálu řeči, ale i kvůli korektnosti výpočtů SNR je ve vývojovém prostředí LabVIEW vytvořen funkční blok *vyhodnocení podmínky pro stejnou délku signálu,* který slouží proto, aby délka dvou zpracovávaných nahrávek, řeči a hluku, měla stejnou délku při pozdějších výpočtech. Jelikož je v programu možnost zvolit si délku signálu a počátek signálu, a všechny nahrávky nemají stejnou délku, tak se tento blok stará o to, aby konečná délka byla dlouhá maximálně jak ta kratší z vybraných. Řešení této funkce lze vidět na obrázku obr. 12:



Obrázek 12: Blok vyhodnocení podmínky pro stejnou délku signálu.

První metoda, která je zpracována v rámci této bakalářské práce, je metoda globálního SNR. Pro zpracování této metody byl použit vzorec (3.13), který je použit v subVI *GSNR* a zobrazen na obrázku:



Obrázek 13: Naimplementování metody GSNR v LabVIEW.

Tato objektivní metoda hodnocení kvality řeči slouží k výpočtu odstupu signálu od šumu přes celý řečový signál, aniž by se v něm nacházela řeč, nebo naopak, že by šlo o výpočet jen v řečově aktivních segmentech. Před samotným výpočtem se na začátku vybere určitá řeč a šum, tyto dvě nahrávky se spolu sečtou a tím nám vyjde zašuměná nahrávka. Z této nahrávky byla dále pomocí

adaptivní filtrace, můžeme vybrat ze čtyř algoritmů LMS, RLS, NLMS a QR-RLS, odfiltrována řeč a šum, které jsou použity pro výpočet globálního SNR. V programu je možná volba amplitudy šumu od 0 do 2násobku původní hodnoty anebo ponechání původní hodnoty k výpočtu. Pomocí této metody jsou spočítány různé hodnoty odstupu signálu od šumu v decibelech, jak je znázorněno v následující tabulce 3.

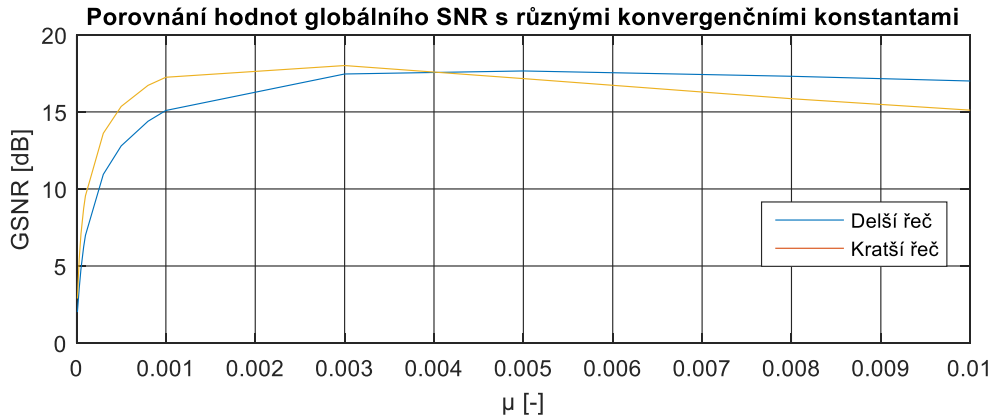
Tabulka 3: Porovnání hodnot globálního SNR s různými konvergenčními konstantami.

Algoritmus	LMS			RLS
Řád filtru (-)	10			10
Parametr adap. algoritmu	μ = 0,0001	μ = 0,001	μ = 0,01	λ = 1
GSNR (dB) Řeč kratší (7 s)	9,5544	17,2486	15,1141	34,15
GSNR (dB) Řeč delší (13 s)	6,989	15,0928	17,0054	31,98

Reč_kratší představuje větu *„Život v lidské společnosti je závislý na schopnosti jednotlivců komunikovat a vzájemně si sdělovat informace.“* A *řeč_delší* zase větu *„Častým problémem detekce řečové aktivity bývá větší množství falešných detekcí řeči, respektive pauz v důsledku prahování kriteriální funkce, která vykazuje krátkodobé fluktuace.“*

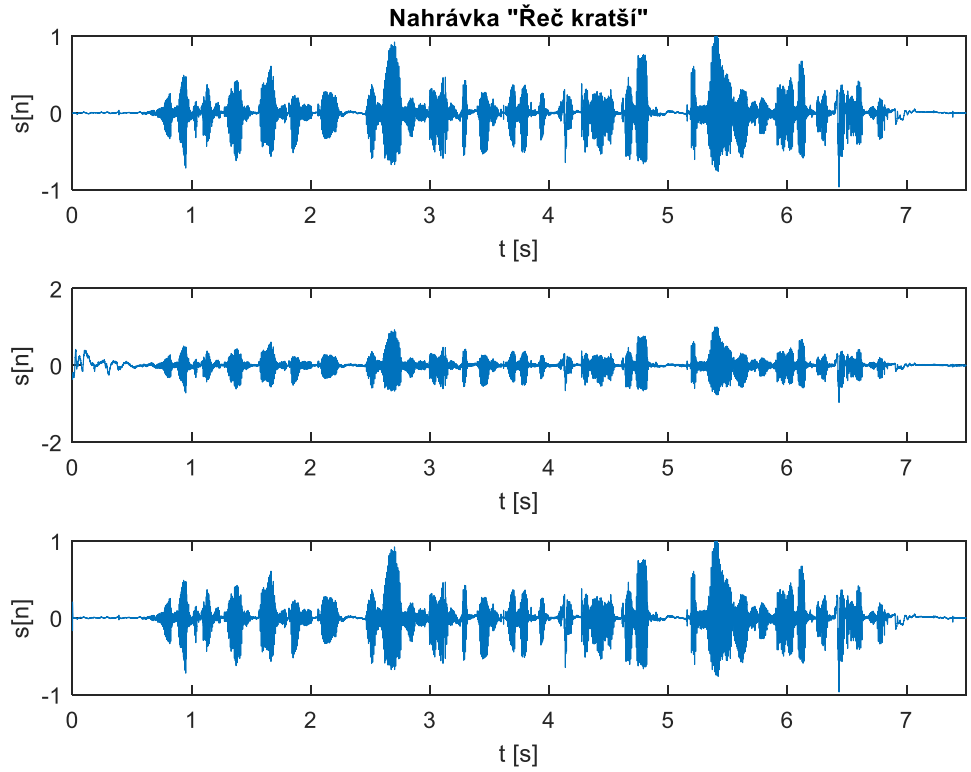
V tabulce 3 μ reprezentuje konvergenční konstantu LMS algoritmu. μ má vliv na rychlost adaptace vah. Jedná se o malou kladnou konstantu, která ovlivňuje vlastnosti adaptace algoritmu. Je-li velká, dochází díky přítomnosti užitečného signálu k rozvažování vah. Pojem malé a velké hodnoty parametru μ souvisí s rozptylem vstupního signálu. V grafu (obr. 14) lze pozorovat vliv velikosti kroku (konvergenční konstanty) na hodnotě globálního odstupu signálu od šumu. K nahrávkám řeči byl přidán hluk křižovatky, který byl následně odfiltrován. Dalším vlivem při zvolení konvergenční konstanty byla i délka nahrávky, kdy pro nahrávku délky 7 sekund byla podle grafu nejlepší hodnota kroku 0,003 a pro 13 sekundovou nahrávku byla optimální hodnota 0,005.

Hodnota λ se používá pro RLS algoritmus a je to faktor zapomínání. Můžeme vidět, že vypočtené hodnoty GSNR, při použití filtrace pomocí RLS algoritmu jsou dvojnásobně větší pro řeč kratší, z čeho vyplývá, že RLS algoritmus dosahuje vyšší kvality filtrace.



Obrázek 14: Porovnání hodnot globálního SNR s různými konvergenčními konstantami.

Při filtraci pomocí LMS algoritmu, je dobře vidět, že tento algoritmus konverguje pomalu. Pokud si přehrajeme tento odfiltrovaný signál, bylo by dobře slyšet, jak postupně klesá hladina hluku. Jde to i dobře vidět na konkrétním průběhu na začátku filtrace (obr. 15). V prvním grafu je zobrazena čistá nahrávka pořízena pomocí měřící karty. Ve druhém grafu, lze na začátku filtrace pozorovat pomalá filtrace LMS algoritmu. Algoritmus tedy potřebuje větší počet iterací, aby se adaptivní filtr přiblížil k optimálnímu stavu filtrace. Naopak filtrace pomocí RLS algoritmu (obr. 15, třetí graf) je velice rychlá, ale za to má mnohem větší nároky na hardware a výpočetní techniku. V praxi jsou nyní požadovány co nejnižší nároky při realizaci adaptivních filtrů, při zachování vysoké kvality potlačení nežádoucího hluku. Z pohledu nákladů na výrobu digitálního zpracování signálů DSP bude výhodnější zkonstruovat algoritmus LMS, ale v budoucnosti můžeme čekat nárůst kvality a výkonu výpočetní techniky, a bude tedy možnost realizovat složitější komplikovanější a výkonnější algoritmy.

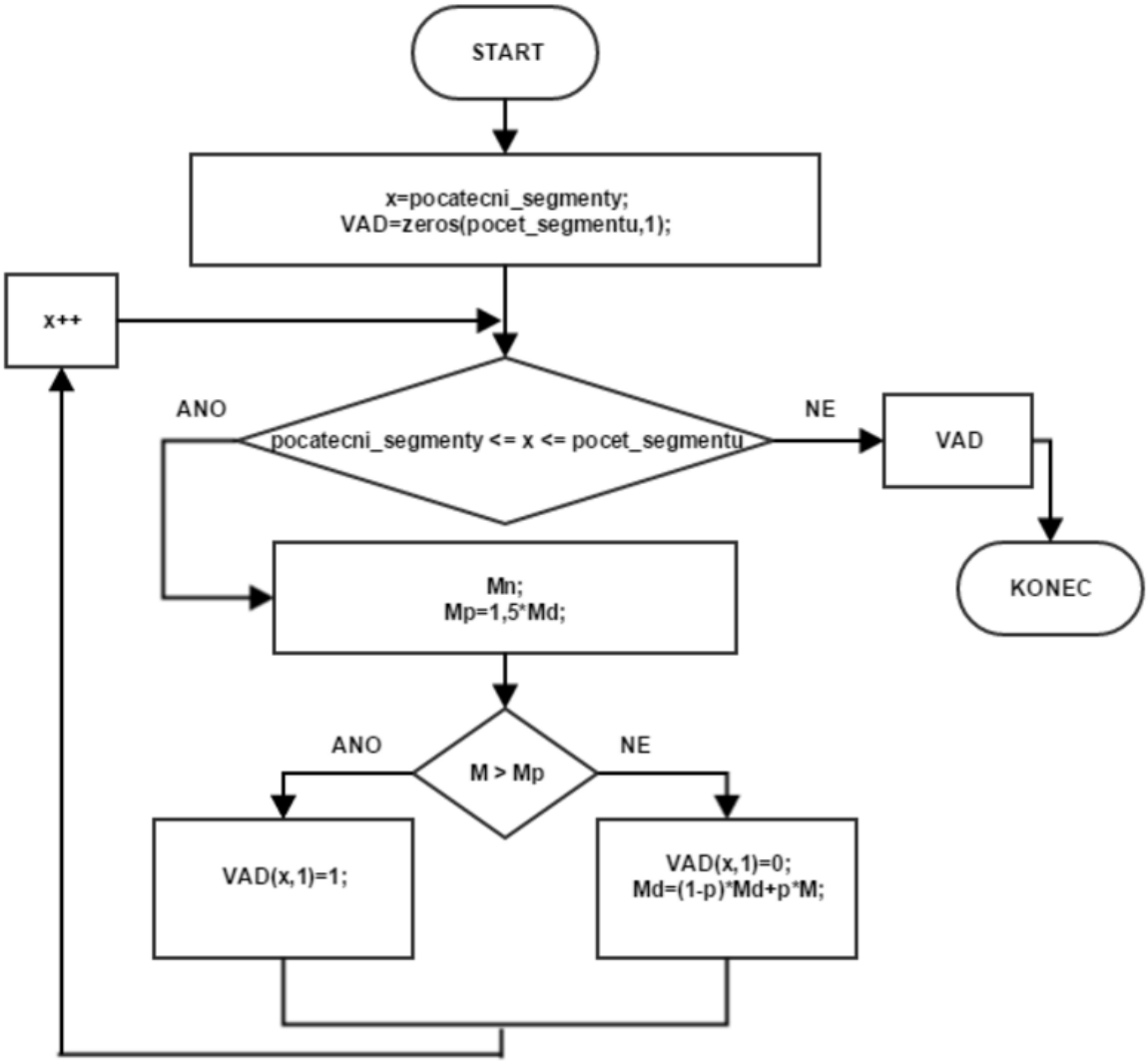


Obrázek 15: Zobrazeni pomalé konvergence LMS algoritmu.

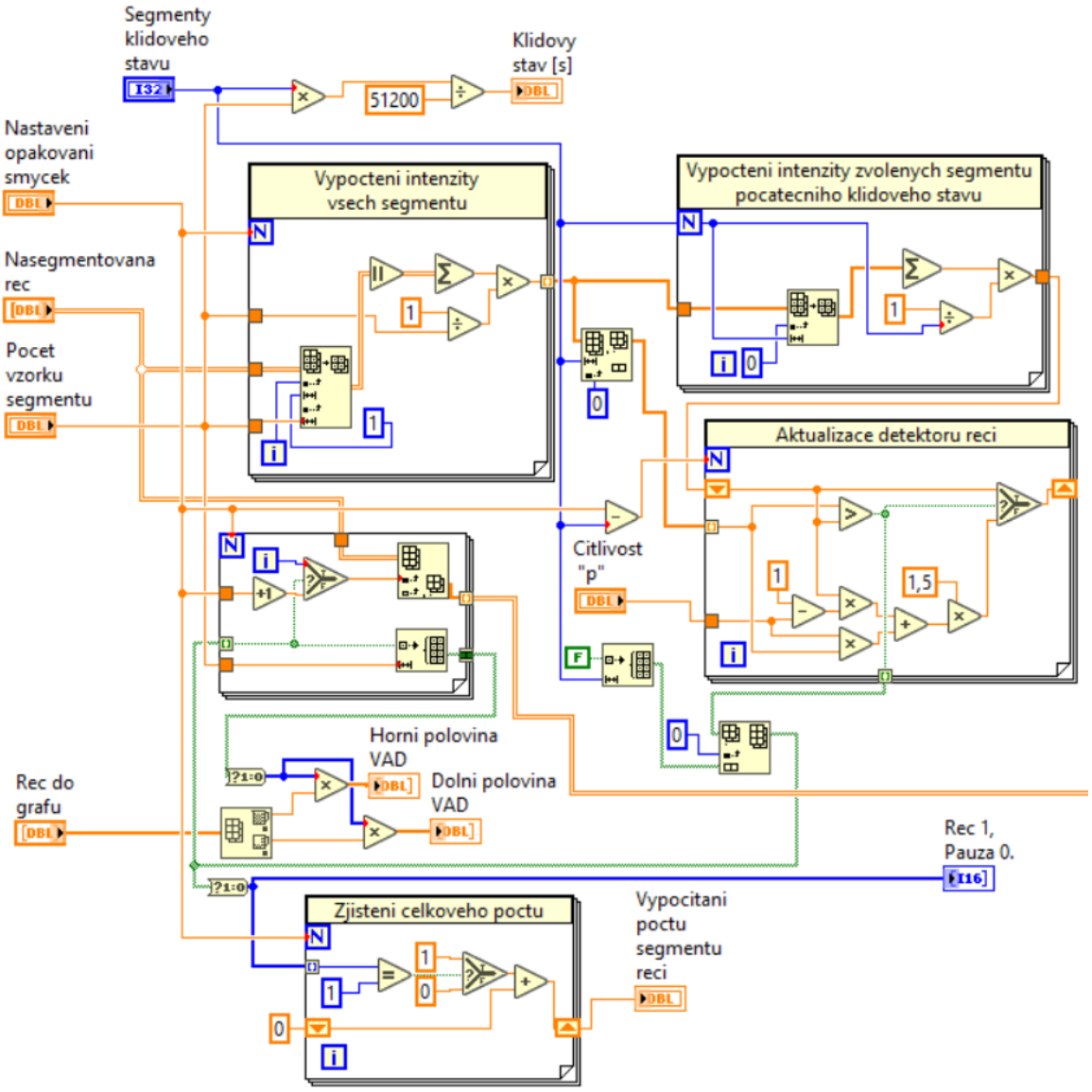
4.1 Zpracování intenzitního detektoru

Pro posuzování další objektivní metody, kterou je segmentální odstup signálu od šumu (SSNR), je důležité použít vhodný detektor řečové aktivit. V této bakalářské práci bude dále popsán intenzitní detektor řeči pro výpočet SSNR. V prvním kroku výpočtu se musí provést segmentace signálu řeči a hluku. Signál lze segmentovat na různě dlouhé rámce v rozmezí od 20 do 32 milisekund, což při vzorkovací frekvenci 51200 Hz znamená, že délky jednotlivých rámců budou velké od 1600 vzorků pro 32 milisekund a 2560 vzorků pro 20 milisekund. Jednotlivé segmenty se ukládají do matice jako sloupcové vektory pro další výpočty. Poté se provede výpočet intenzity pro každý segment podle vzorce (3.18) a hodnoty intenzity se uloží do sloupcového vektoru. Po výpočtu intenzit si uživatel zvolí počet inicializačních segmentů x, ze kterých se spočítá podle vzorce (3.19) intenzita hluku počátečních segmentů M_d, která bude sloužit jako podklad pro určení detekčního prahu M_p. Po provedení inicializace detekčního prahu může být provedeno detekování řeči v signálu, kdy se bude porovnávat intenzita testovaného segmentu s prahovou intenzitou M_p vypočtenou podle vztahu (3.20). Cyklus se provádí pro segmenty začínající za segmenty inicializačními (*pocatecni_segmenty*) až do konečného počtu segmentů (*počet_segmentu*). Dále se budou hodnoty intenzit jednotlivých segmentů porovnávat s prahovou hodnotou intenzity. V případě, že bude intenzita segmentu menší, než je prahová hodnota dojde k aktualizaci prahové hodnoty podle vzorce (3.21). V tomto případě je porovnávaný segment řečově neaktivní a do proměnné *VAD,*

která slouží pro zapsání detekovaných segmentů, je na aktuální pozici zapsána hodnota 0. V případě detekce řečové aktivity se prahová hodnota intenzity aktualizovat nebude, její hodnota bude pouze přenesena do dalšího cyklu a do proměnné *VAD* se zapíše hodnota 1. Jednotlivé segmenty se ukládají jako sloupcový vektor, které slouží k dalšímu zpracování. Na obrázku (Obr. 16) je zobrazen vývojový diagram průběhu detekce a aktualizace prahové hodnoty intenzity.



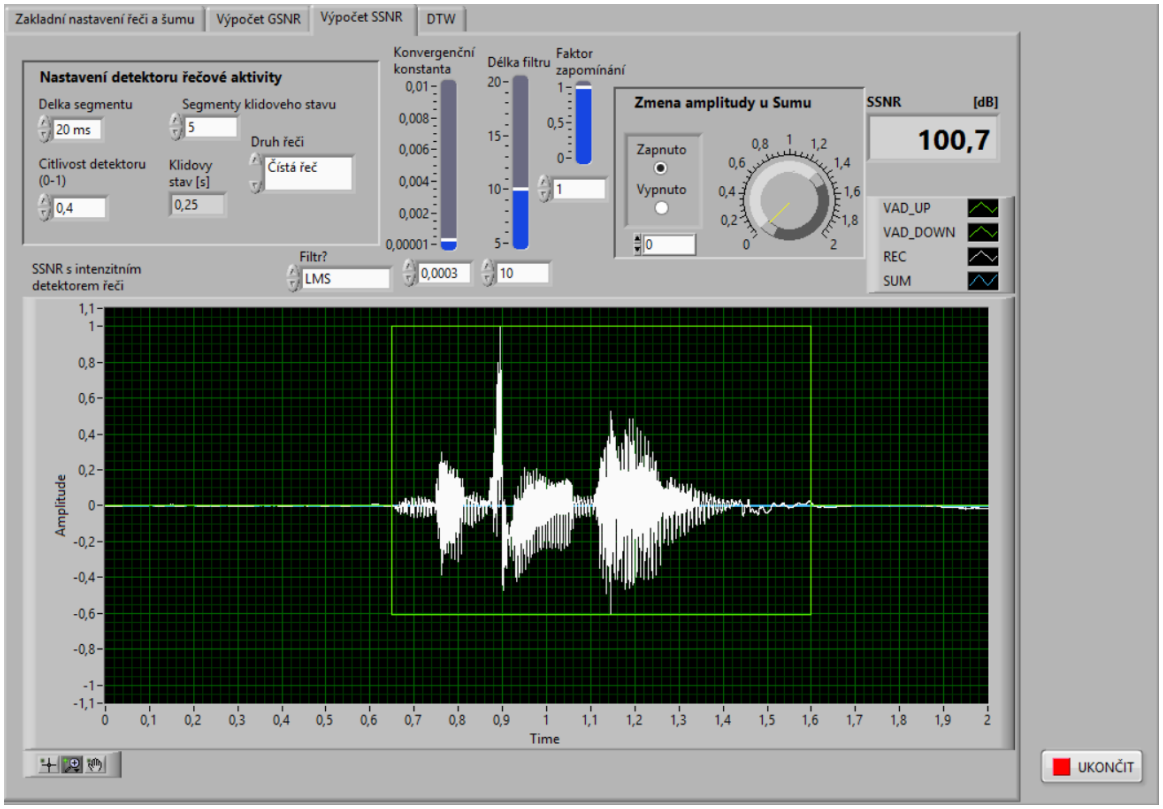
Obrázek 16: Vývojový diagram průběhu detekce intenzitního detektoru.



Obrázek 17: Implementace detektoru řečové aktivity v LabVIEW.

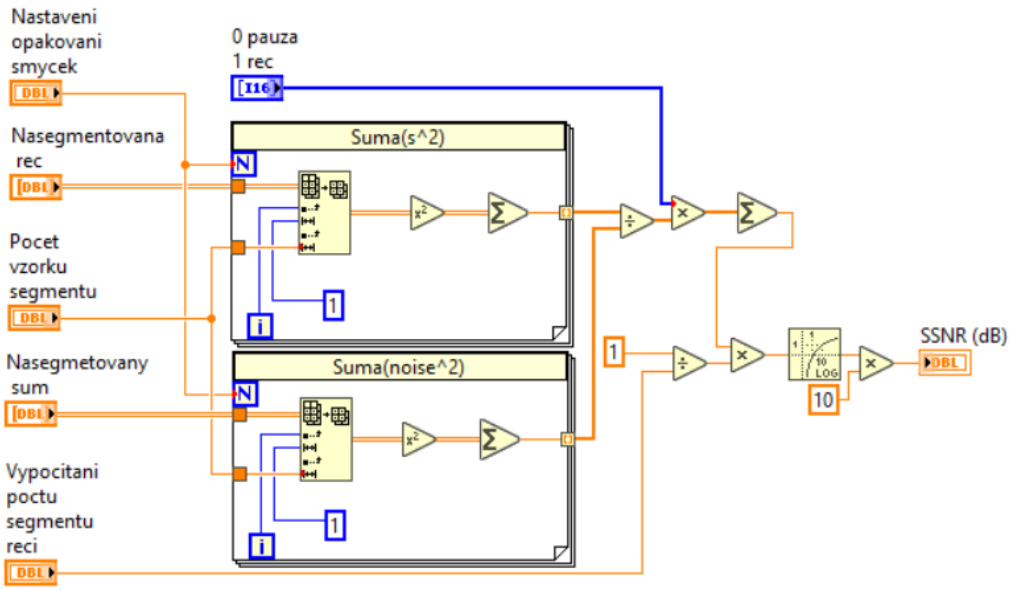
Předchozí obrázek zobrazuje implementaci detektoru řečové aktivity v prostředí LabVIEW. Ve smyčce *Vypocteni intenzity vsech segmentu* je proveden výpočet intenzit podle vzorce (3.18). Jako následující krok se provede *Vypocet intenzity zvolenych segmentu pocatecniho klidového stavu* pro zvoleny počet klidových segmentu. Tyto klidové segmenty jsou dále vynechány a výpočet *Aktualizace detektoru reci* se bude provádět až pro následující segmenty. Během aktualizace detektoru se zároveň vyhodnocují hodnoty detektoru řečové aktivity, ze kterých se následně vypočítá celkový počet segmentů řeči a průběh detekce řečově aktivních segmentů.

Jakmile jsou propočítány všechny segmenty a uložené v proměnné *VAD*, kde se vyskytují segmenty řečově aktivní, lze provést výpočet segmentálního odstupu signálu od šumu (dále jen SSNR). Intenzitní a energetické detektory, jsou vhodné zejména tam, kde je hluk spíš stacionární, tj. neměnící se s časem, jelikož jsou tyto detektory citlivé na dynamické změny rušivého prostředí. Obrázek 18 znázorňuje detekci dle vzorce (3.16) slova „*Dobrý den*“ bez přidaného hluku.



Obrázek 18: Nahrávka "Dobrý den" s detekcí řeči.

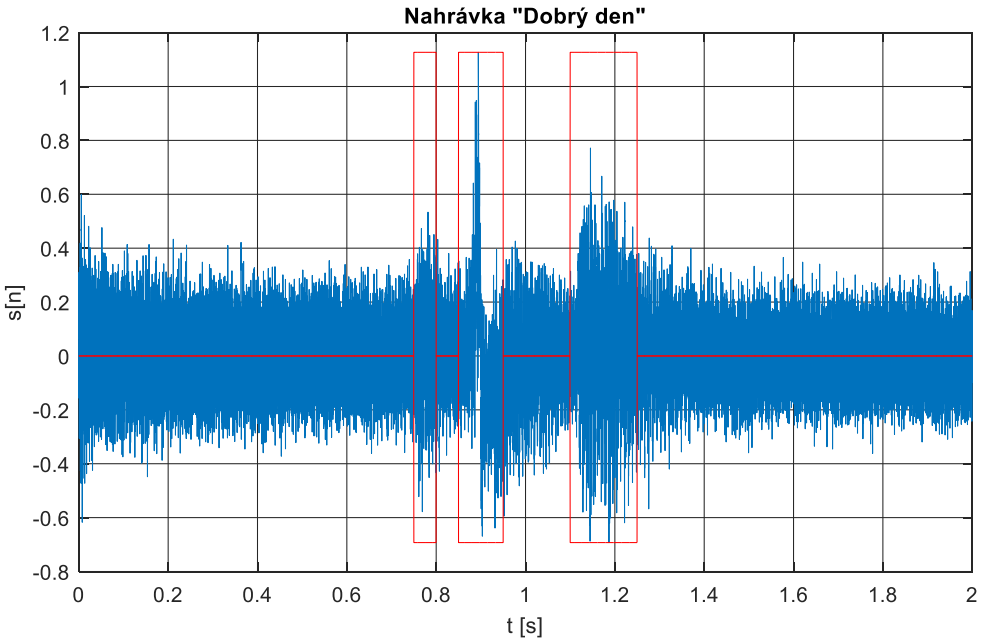
Na obrázku je zobrazeno uživatelské rozhraní, sloužící pro výpočet SSNR. V *Nastavení detektoru řečové aktivity* si můžeme zvolit délku segmentů, počet segmentů, které budou složit jako inicializační, *Klidovy stav* zobrazuje čas inicializace v sekundách, a dále citlivost detektoru, která se obvykle volí v rozmezí od 0,4 do 0,6.



Obrázek 19: Naimplementování metody SSNR v LabVIEW.

Předchozí obrázek zobrazuje výpočet segmentálního odstupu signálu od šumu v prostředí LabVIEW. Výpočet je proveden podle vzorce (3.16), proměnná VAD_i představuje hodnotu 0 pro pauzu a 1 pro řeč. Hodnota K_{VAD} představuje celkový počet segmentů řeči detekovaný detektorem v dané nahrávce.

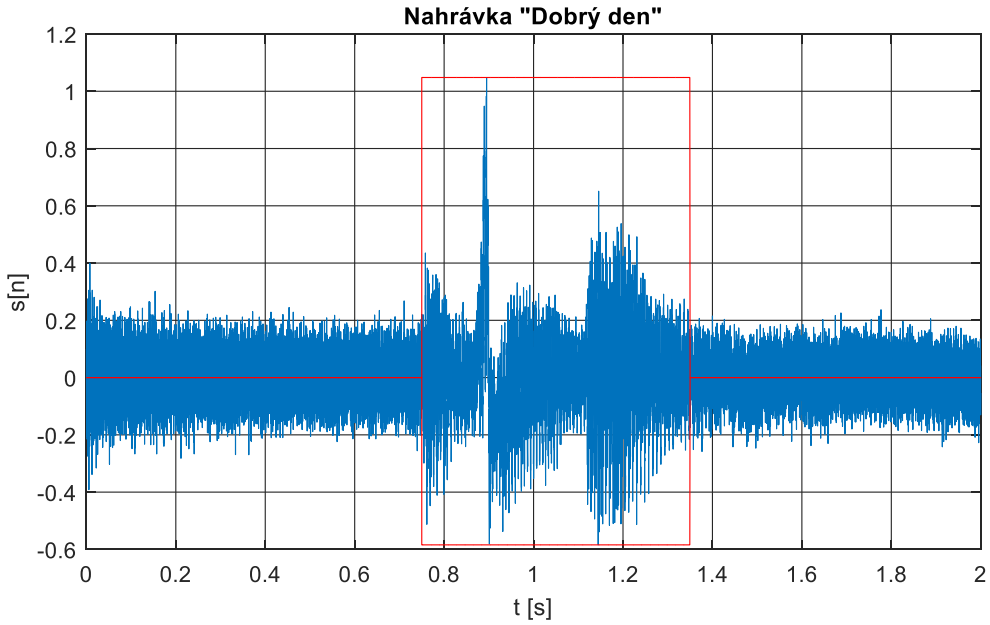
Graf (obr. 18) zobrazuje, že detekce z čistého signálu řeči, který není nijak ovlivněn šumem, proběhla výborně. Jakmile se k nahrávce přidá šum, a sníží se GSNR přibližně pod úroveň 4 dB, detektor již nedokáže pořádně rozeznat začátek, konec a ani průběh slova. Je to způsobeno z důvodu, že segmenty jsou překryty stacionárním šumem, a proto u segmentů, ve kterých se vyskytují nižší úrovně intenzity, může dojít k chybě detekce řeči. Na obrázku (obr. 20) je vidět nepovedená detekce řeči kdy byla vypočtená hodnota GSNR na 2 dB.



Obrázek 20: Detekce řeči při hodnotě GSNR = 2 dB.

Až při hodnotě GSNR větší, než jsou 4 dB vyšla detekce slova téměř stejně jako při detekci slova čistého signálu. Nepřesnost je způsobena z důvodu, že detekce slova proběhla v rozmezí 0,75 s až 1,35 s, přičemž detekce čistého slova byla provedena v úseku 0,65 s až 1,5 s, jak lze pozorovat na obrázku 21.

Z těchto závěrů tedy plyne, že detekce řeči, která se provádí v zašuměném prostředí je problematická, zvlášť pak u detektorů, které pracují na výpočtech intenzit nebo energie a jsou tedy počítány v časové oblasti. Kdyby tedy došlo k tomu, že by rušivý signál obsahoval velké výkyvy intenzity, tak by detektor nedokázal rozeznat, zda se jedná o úseky řeči nebo pauzy, jelikož hodnoty intenzity by byly podobné kvůli výkyvům. Z tohoto důvodu byly pořízeny nahrávky, které by neměly obsahovat téměř žádný hluk, a tudíž mít velký poměr *SNR*. Dále bylo použito adaptivních filtrů, aby při spojení řeči s hlukem nedocházelo při odfiltrování ke špatné detekci řeči.



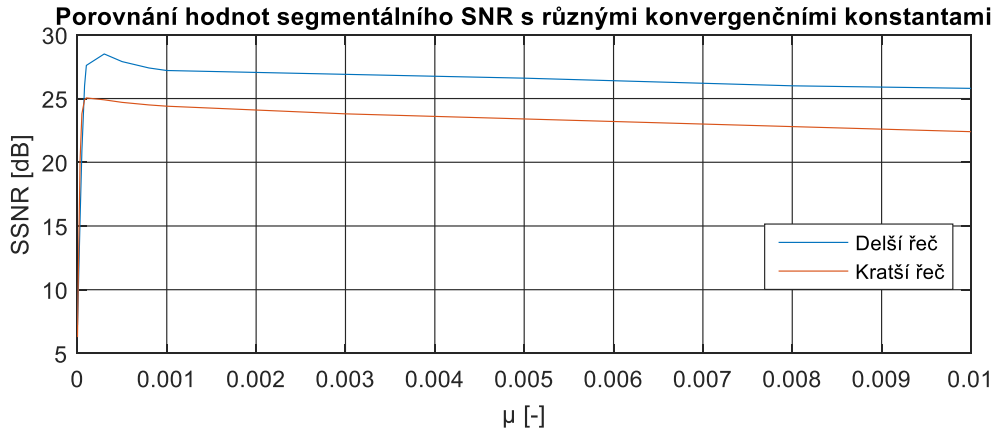
Obrázek 21: Detekce řeči při hodnotě GSNR = 4 dB.

Pomocí detektoru řeči byly spočítány hodnoty SSNR pro různé konvergenční konstanty, kdy byly použity stejné nahrávky řeči a k nim připojen hluk křižovatky. Jako algoritmus pro adaptivní filtr byl použit LMS a RLS. Výsledek je zobrazen v tabulce. Hlavním rozdílem, který mezi těmito dvěma metodami nastal, je ten, že hodnoty SSNR oproti GSNR jsou větší o více než 10 dB pro algoritmus LMS. Při použití algoritmu RLS se hodnoty pohybují kolem 40 dB.

V grafu (obr. 22) můžeme opět vidět vliv velikosti kroku (konvergenční konstanty) na hodnotu segmentálního odstupu signálu od šumu. Pro *řeč_delší* vyšla nejvyšší hodnota SSNR 28,5 dB při konvergenční konstantě 0,0003 a pro *řeč_kratší* byla nejvyšší hodnota SSNR rovna 25,05 dB při velikosti kroku 0,0001.

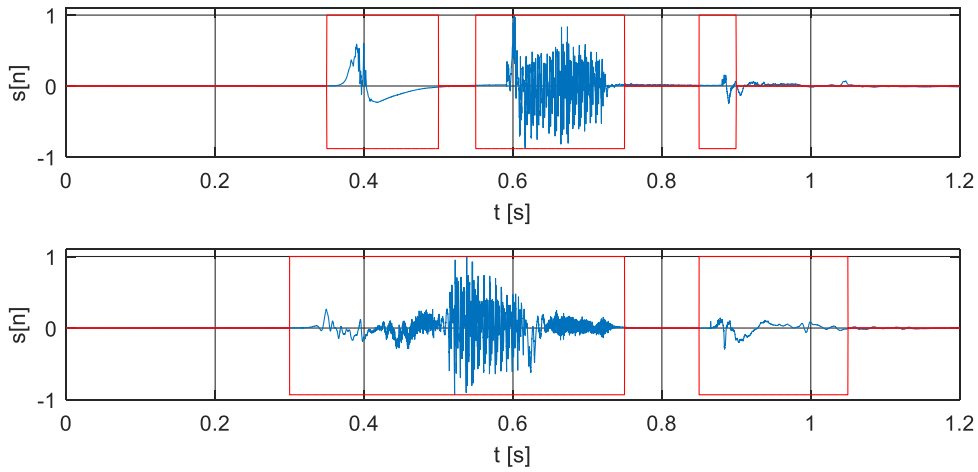
Tabulka 4: Porovnání hodnot segmentálního SNR s různými konvergenčními konstantami.

Algoritmus	LMS			RLS
Řád filtru (-)	10			10
Parametr adap. algoritmu	μ = 0,0001	μ = 0,001	μ = 0,01	λ = 1
SSNR (dB) Řeč kratší (7s)	25,05	24,46	22,45	39,38
SSNR (dB) Řeč delší (13s)	27,64	27,21	25,82	38,94



Obrázek 22: Porovnání hodnot segmentálního SNR s různými konvergenčními konstantami.

Poslední stěžejní metodou, která byla v rámci této bakalářské práce zpracována, byla subjektivní metoda dynamického borcení času (Dynamic Time Warping), která slouží pro rozpoznávání izolovaných slov. Základem této metody bylo vypočítat PARCOR koeficienty. Tyto koeficienty jsou počítány pouze pro řečově aktivní segmenty a je zde velice důležitá kvalita detekce, kdy problém nastává u slov, kde se přibližně uprostřed slova nachází okluzíva, tedy ve slově se nachází krátká pauza o velikosti přibližně 10 milisekund. Tato pauza vede k pořízení segmentu, ve kterém se nevyskytuje řeč uprostřed slova a tím ke špatné detekci. Případ je zobrazen na obrázku 23:

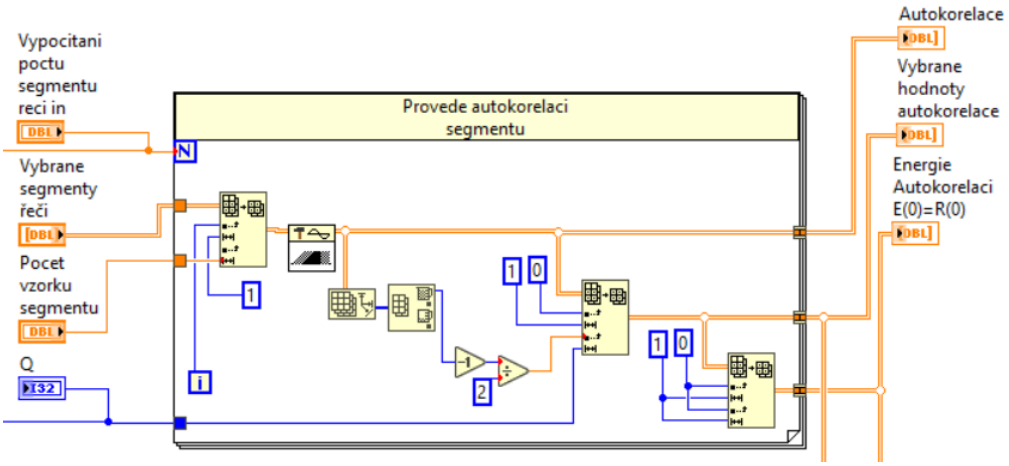


Obrázek 23: Případ špatné detekce slova.

V horním grafu je zobrazena nahrávka slova *„pět“*, která za písmenem *p* a před písmenem *t* má pauzu, tato pauza vede ke špatné detekci detektoru, a to stejné lze vidět u nahrávky *„šest“* taktéž před písmenem *t*.

Pro výpočet PARCOR koeficientů je třeba znát řád predikce, který je vypočítán podle vzorce (2.16) a je vyčíslen na hodnotu 55 pro tuto bakalářskou práci. Z důvodů pořízení referenčních nahrávek při délce segmentu 20 ms je důležité použít tuto délku segmentů i pro nahrávky testované. Prvním krokem při výpočtu PARCOR koeficientů je, že se musí vypočítat autokorelační koeficienty

použitím funkce *AutoCorrelation.vi*. Implementace v prostředí LabVIEW je znázorněna na obrázku:

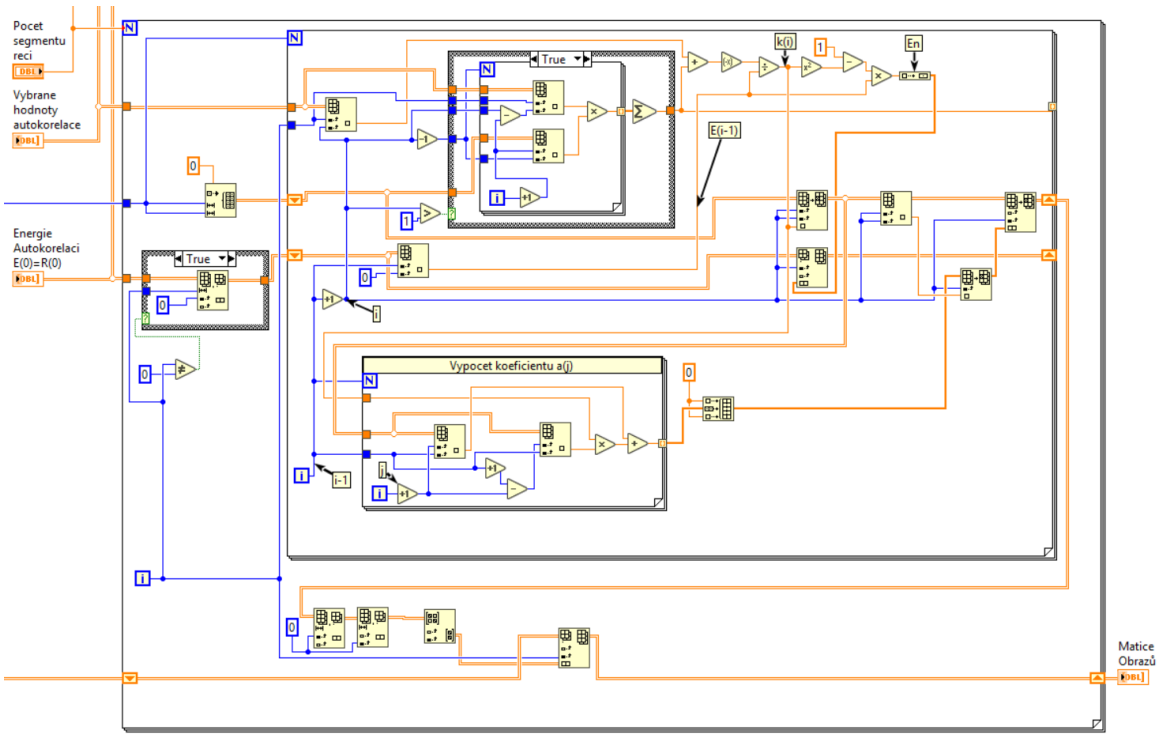


Obrázek 24: Implementace pro výpočet autokorelačních koeficientů v prostředí LabVIEW.

Dále bylo důležité vybrat 55 hodnot autokorelačních koeficientů od $R(0)$ do $R(Q)$:

$$index = \frac{(size(\boldsymbol{R}) - 1)}{2}. \tag{4.1}$$

Dále je třeba si pro každý řečově aktivní segment uložit těchto 55 hodnot do řádkové matice, jejíž počet řádků bude roven počtu řečově aktivních segmentů. Z této matice se potom aplikací algoritmu Levinson-Durbina (2.15) získá třírozměrná matice, která bude mít tvar, že pod diagonálou budou všechny prvky rovny hodnotě 0, jak je znázorněno v matici (2.17). Poté je třeba získat z každé matice právě PARCOR koeficienty testovaného slova, které se uloží do matice *PARCOR_B*, kde bude počet segmentů řeči znázorňovat počet sloupců v matici. Hodnoty *PARCOR_A* jsou stejné koeficienty spočtené pro referenční nahrávky. Obrázek 25 zobrazuje výpočet matic, pro všechny řečově aktivní segmenty, podle Levinson-Durbinova algoritmu, ze kterých jsou následně vybrány koeficienty odrazu k_i neboli PARCOR koeficienty v prostředí LabVIEW. Tyto koeficienty jsou uloženy jako sloupcové vektory do *Matice obrazů*.



Obrázek 25: Implementace výpočtů PARCOR koeficientů v prostředí LabVIEW.

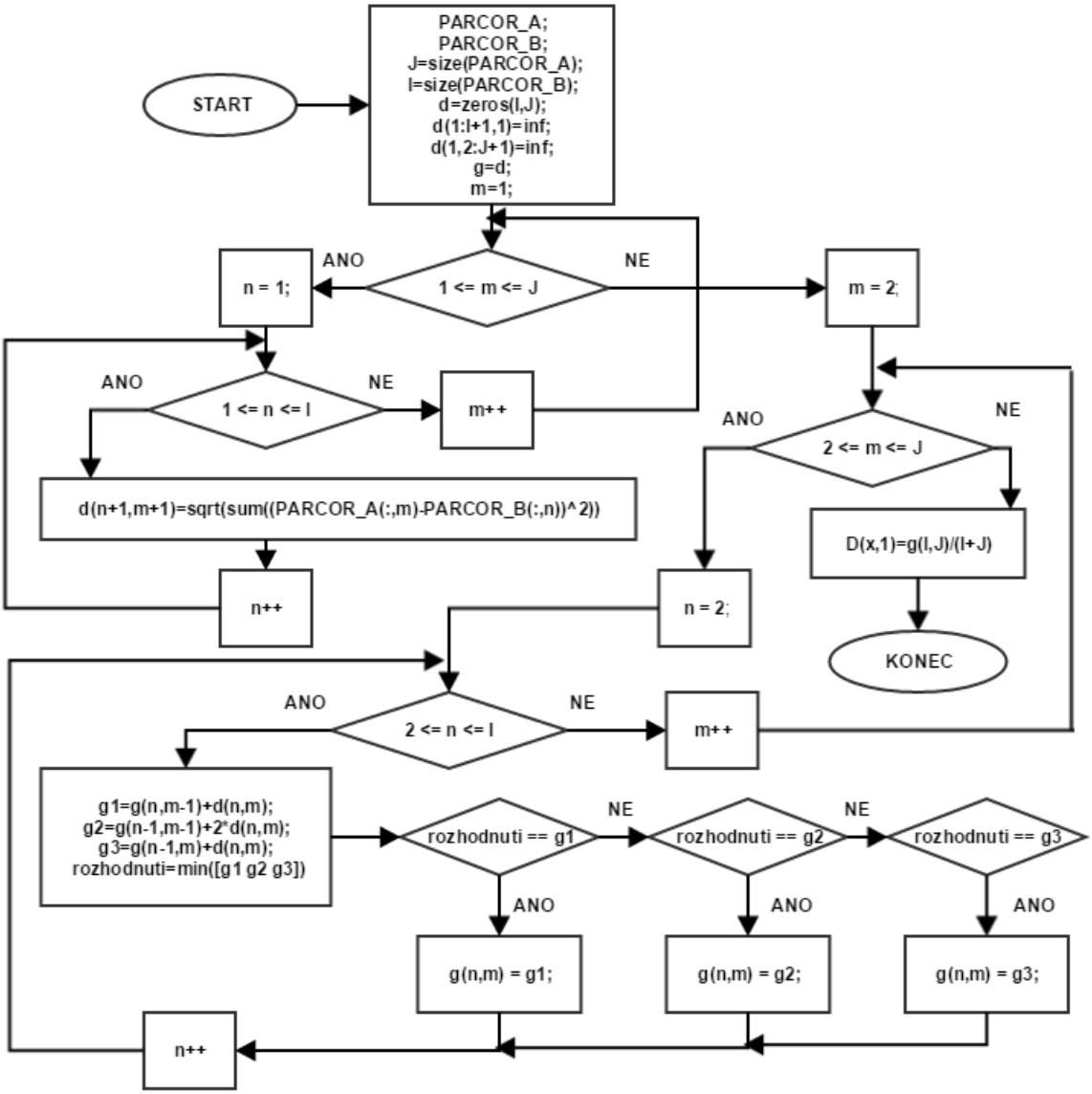
Po dokončení výpočtů koeficientů můžeme popsat algoritmus na obrázku (Obr. 26) pro výpočet metody DTW. V prvním kroku budou načteny obě matice *PARCOR_A* a *PARCOR_B*, které reprezentují slovo referenční a testované. V dalším kroku dojde k vytvoření matice lokálních vzdáleností *d* jako nulové, kromě prvního řádku od druhé hodnoty a sloupce od druhé hodnoty, kde budou vloženy hodnoty *inf*. Je to z toho důvodu, aby při pozdějších výpočtech byl přesně určen počáteční krok při vzdálenosti *D*, tedy muselo dojít k časovému posunu vpřed. V každém cyklu výpočtu vzdáleností, mezi *m*-tým až *J*-tým vektorem referenčního slova a mezi *n*-tým až *I*-tým vektorem testovaného slova, je vypočtení vzájemné lokální vzdálenosti *d*, podle vzorce využívajícího euklidovskou míru:

$$d(n+1,m+1) = \left[\sum_{i=1}^{Q} |PARCOR\_A(i,m) - PARCOR\_B(i,n)|^2\right]^{\frac{1}{2}} \tag{4.2}$$

Dalším krokem je výpočet akumulovaných vzdáleností *g* podle rovnice (3.8), kde se vytvoří stejnojmenná matice *g*. Hledá se nejmenší vypočtená hodnota podle vzorce, která se uloží na konkrétní místo v matici podle aktuálních hodnot *n* a *m*. Celková vzdálenost mezi testovanou sekvencí a referenční je vypočtena podle vzorce:

$$D(x,1) = g(I,J)/(I+J) \tag{4.3}$$

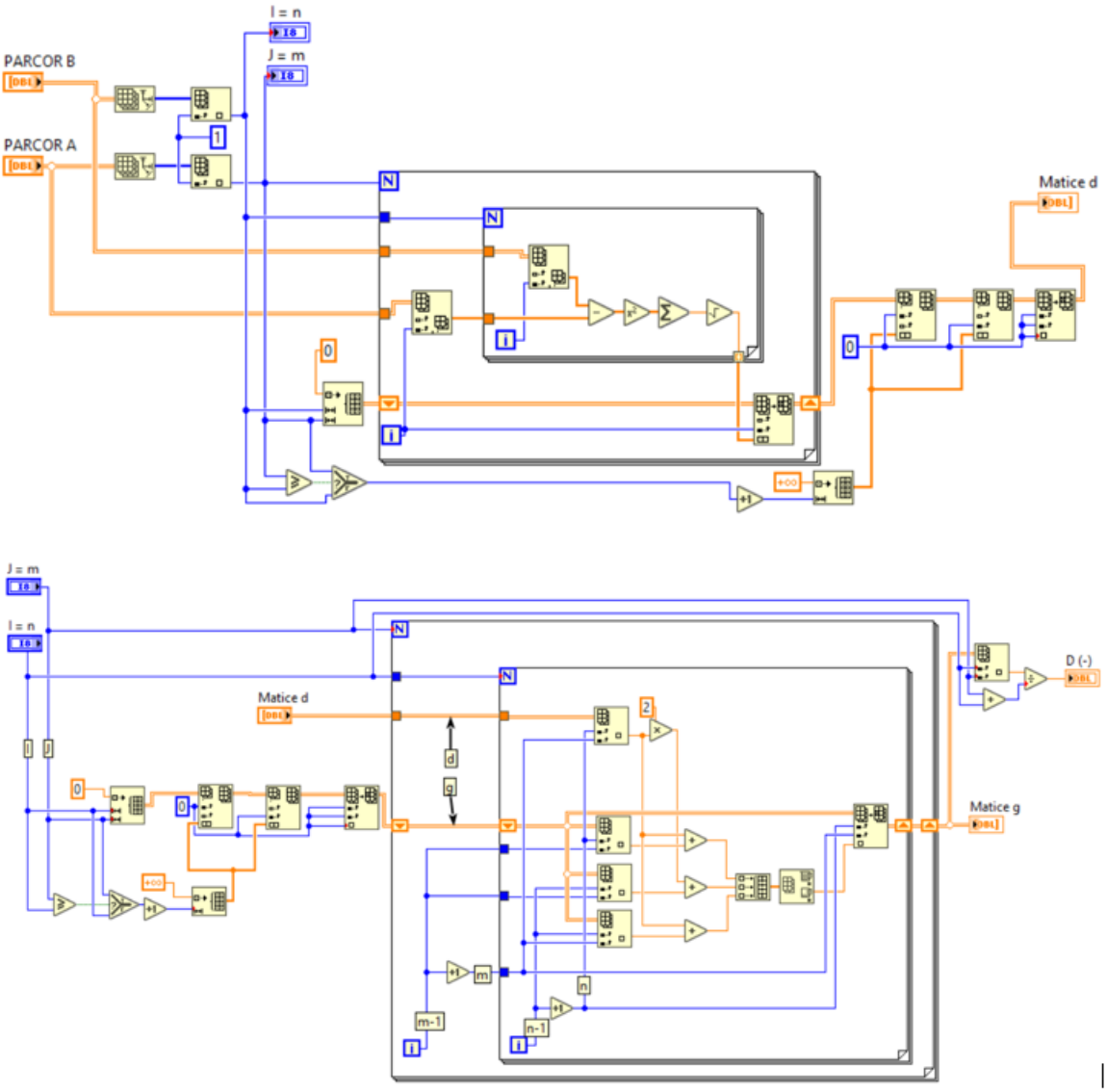
Ve výsledném programu je zobrazena vzdálenost *D* jak v rozmezí od <0 1>, kde nižší číslo nám říká, že čísla jsou si podobnější, tak je podobnost přepočítána a zobrazena do procentuální vzdálenosti, kde nám hodnota 100 % zajišťuje největší podobnost.



Obrázek 26: Blokový diagram subjektivní metody DTW.

Pro samotné rozpoznávání izolovaných slov pomocí metody DTW byly pořízeny nahrávky slov *jedna, dva, tři, čtyři, pět, šest, sedm, osm, devět*, které byly nahrány jedním řečníkem. Z těchto slov byla vytvořena knihovna, z níž si uživatel načte konkrétní slovo, během výpočtů si může uživatel neustále měnit vybrané slovo z knihovny, a toto slovo bude považováno za referenční. Dále si pomocí intenzitního detektoru z metody SSNR vybere slovo, které se bude považovat za slovo testované.

Na obrázku 27 je v první části zobrazena implementace pro výpočet lokálních vzdáleností *d* mezi dvěma porovnávanými slovy. Druhá část zobrazuje výpočet akumulovaných vzdáleností *g* podle rovnice (3.8), kde je zobrazen výběr nejmenší hodnoty *g*, která je následně uložena na konkrétní místo matice. Po dokončení výpočtu matice akumulovaných vzdáleností proběhne výpočet konečné vzdálenosti *D* mezi porovnávanými slovy podle vzorce (4.3).



Obrázek 27: Implementace metody DTW v prostředí LabVIEW.

V rámci této bakalářské práce jsem porovnával referenční nahrávky, tj. nahrávky bez hluku, s nahrávkami znovu detekovanými detektorem a referenční nahrávky s nahrávkami řeči po filtraci, a tím byla dobře vidět i kvalita filtrace.

V následujících tabulkách (Tab. 5, tab. 6) je zobrazeno porovnání slov „*jedna*“ a „*čtyři*“ s ostatními slovy „*jedna*“ až „*devět*“ a můžeme sledovat, že i když byla porovnávána stejná slova bez jakékoliv filtrace, nemají slova stoprocentní podobnost.

Tabulka 5: Vypočítané hodnoty vzdáleností D porovnáním referenčního slova „jedna“ se slovy "dva" až "devět".

	Jedna Jedna	*Jedna Dva*	*Jedna Tři*	*Jedna Čtyři*	*Jedna Pět*	*Jedna Šest*	*Jedna Sedm*	*Jedna Osm*	*Jedna Devět*
D (-)	0,0095	0,5344	0,599	0,6741	0,5661	0,6375	0,7651	0,5838	0,5551
D (%)	99,05	46,56	40,03	32,59	43,39	36,35	23,49	41,62	44,49

Tabulka 6: Vypočítané hodnoty vzdáleností D porovnáním referenčního slova "čtyři" se slovy "jedna", "dva", "tři", "pět" až "devět".

	Čtyři Jedna	*Čtyři Dva*	*Čtyři Tři*	*Čtyři Čtyři*	*Čtyři Pět*	*Čtyři Šest*	*Čtyři Sedm*	*Čtyři Osm*	*Čtyři Devět*
D (-)	0,6742	0,6017	0,5284	0,0007	0,5043	0,5737	0,7012	0,6351	0,5085
D (%)	32,58	39,83	47,16	99,93	49,57	42,63	29,88	36,49	49,15

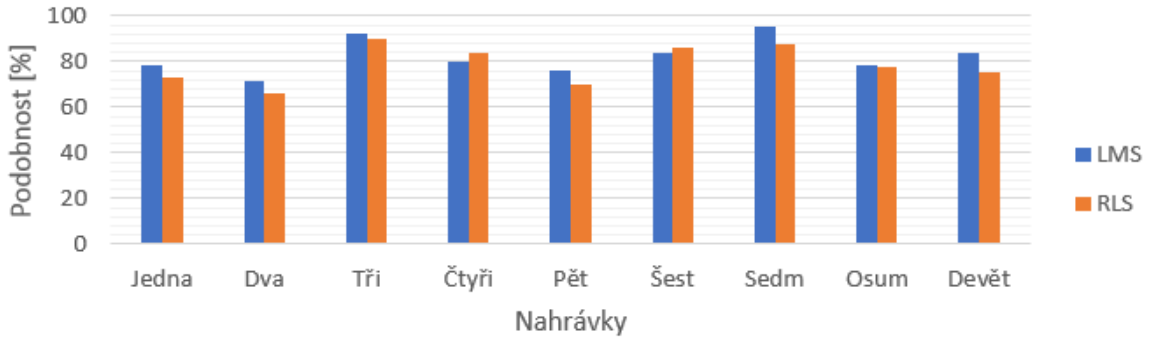
V následující tabulce (tab. 7) je zobrazeno porovnání jednotlivých slov se slovy, která byla odfiltrována použitím adaptivní filtrace pro algoritmus LMS, který měl nastavenou konvergenční konstantu $\mu = 0{,}0003$ a řád filtru byl nastaven na hodnotu 5. V druhé části tabulky je porovnána reference spolu s odfiltrovanou řečí, pomocí algoritmu RLS, který měl nastavený faktor zapomínání na 1.

Pomocí metody DTW vyšla podobnost slov referenčních více než 99 % ve všech testech. Při posuzování podobnosti reference s odfiltrovanou řečí bylo použito dvou algoritmů adaptivních filtrů. První algoritmus byl použit LMS, u kterého vycházela podobnost od 75 % až do 95 %. Druhým algoritmem pro posuzování podobnosti byl použit RLS algoritmus. Jeho hodnoty podobnosti se v procentuální části vyskytovali téměř shodné jako u LMS. U hodnot SSNR vyšel lépe algoritmus RLS, jehož hodnoty vycházeli větší i o 20 dB.

Tabulka 7: Výpočet vzdáleností D s referenční nahrávkou odfiltrovanou řečí.

	Odfiltrovaná řeč (LMS)			Odfiltrovaná řeč (RLS)		
	D (-)	D (%)	SSNR (dB)	D (-)	D (%)	SSNR (dB)
Jedna	0,2193	78,07 %	23,84	0,2722	72,78 %	41,40
Dva	0,2827	71,73 %	25,14	0,3416	65,84 %	36,86
Tři	0,0816	91,84 %	28,29	0,1042	89,58 %	39,79
Čtyři	0,2036	79,64 %	25,89	0,1631	83,69 %	40,41
Pět	0,2421	75,79 %	21,12	0,3027	69,73 %	36,39
Šest	0,1597	84,03 %	20,06	0,1416	85,84 %	36,34
Sedm	0,0511	94,89 %	27,03	0,1220	87,80 %	41,72
Osm	0,2145	78,55 %	23,96	0,2260	77,40 %	38,14
Devět	0,1649	83,51 %	28,62	0,2498	75,02 %	40,69

Další tabulka ukazuje porovnání čtyř algoritmů pro adaptivní filtry. Lze vidět, že nejlepších výsledků dosahoval algoritmus RLS. Hodnota faktoru zapomínání, která se u tohoto filtru nastavuje místo konvergenční konstanty, byla nastavena na $\lambda = 1$, tedy má nekonečnou paměť. Jak je zobrazeno na obrázku 29, u RLS algoritmu se nevyskytuje na počátku filtrace téměř žádná hladina hluku, kterou je možné pozorovat u LMS algoritmu, algoritmus tedy konverguje mnohem rychleji než LMS. Podobnosti nahrávek vychází podobně i pro další dva algoritmy NLMS a QR-RLS.

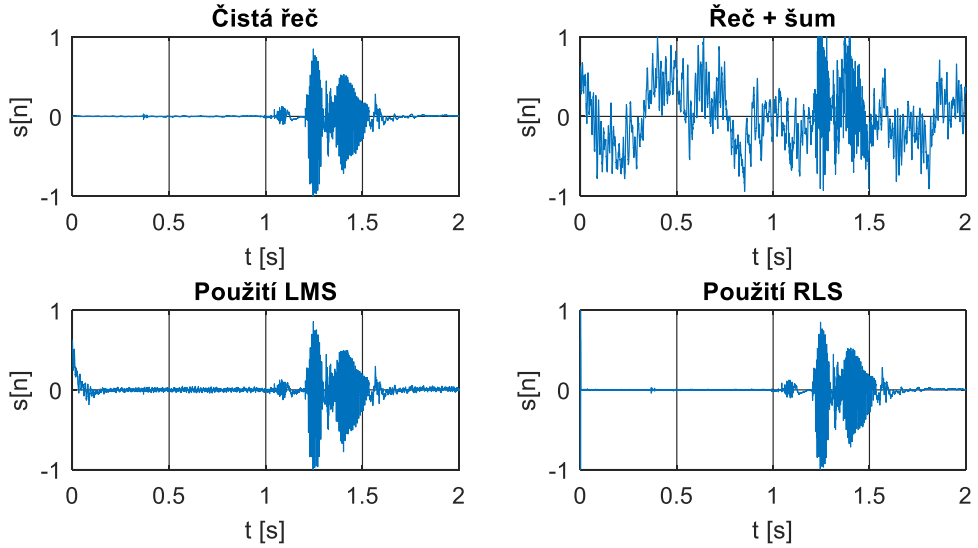


Obrázek 28: Histogram zobrazení podobnosti pro LMS a RLS.

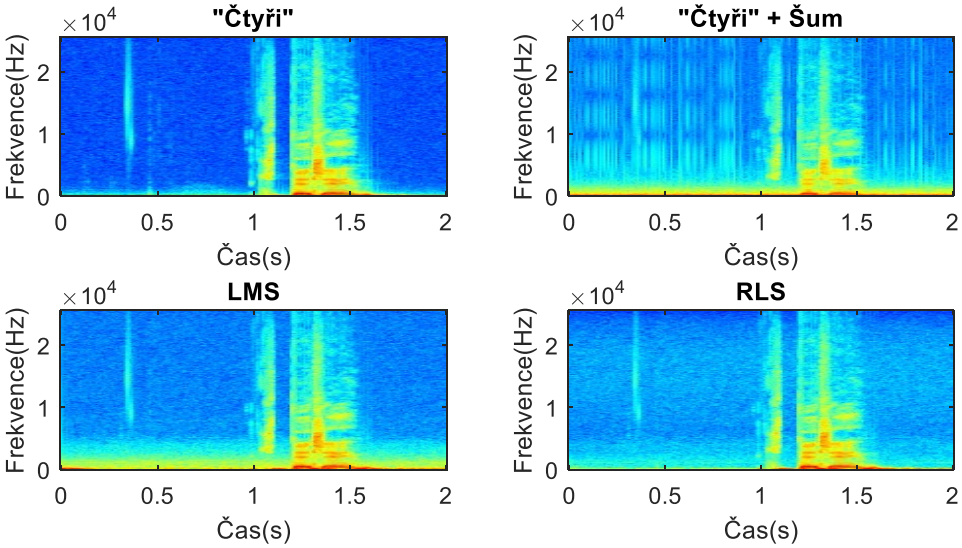
Tabulka 8: Porovnání nahrávky "čtyři" s různými algoritmy.

Nahrávka	Čtyři		
Algoritmus	**D (-)**	**D (%)**	**SSNR (dB)**
LMS	0,2036	79,64	20,49
NLMS	0,2495	75,05	15,38
RLS	0,1631	83,69	40,41
QR-RLS	0,1632	83,68	40,41

Na obrázcích 29 a 30 lze pozorovat úspěšnost filtrace. Nejlepších výsledků podle hodnot SSNR dosahoval algoritmus RLS. Z grafu lze pozorovat, že řeč odfiltrovaná tímto algoritmem se nejvíce podobá čisté řečové nahrávce. Algoritmus RLS, který má nejlepší přesnost a rychlost filtrace, má i přes to nevýhodu a tou je matematická náročnost této filtrace.



Obrázek 29: Porovnání adaptivních filtrů v časové oblasti.



Obrázek 30: Porovnání adaptivních filtrů ve spektrogramu.

Závěr

Cílem této bakalářské práce byla implementace objektivních a subjektivních metod pro posuzování kvality řeči v programovacím a vývojovém prostředí LabVIEW. Z objektivních metod byly vybrány základní metody odstupu signálu od šumu GSNR a segmentální odstup signálu od šumu SSNR. Velký rozdíl mezi výsledky metody globálního SNR a segmentálního SNR s využitím VAD je způsoben tím, že VAD vyhodnocuje segmenty pomocí krátkodobé intenzity a pro silněji zarušené nahrávky je krátkodobá intenzita hluku vyšší než krátkodobá intenzita neznělých hlásek. Potom se tyto segmenty vyhodnotí jako šumové a tím se snižuje výsledný poměr signálu od šumu. V rámci bakalářské práci byl použit detektor řeči, který využívá výpočtu intenzit jednotlivých segmentů. Výhodou intenzitního detektoru řečové aktivity je jeho jednoduchost a výpočetní nenáročnost. Jako subjektivní metoda bylo zvoleno dynamické borcení časové osy (DTW). Metoda DTW byla počítána pomocí PARCOR koeficientů, podle Durbin-Levinsonova algoritmu popsaného teoretické části. Metodou DTW je vyhodnocena kromě podobnosti slov mezi referenčními slovy taky slova, která byla odfiltrována pomocí adaptivních filtrů, využitím algoritmů LMS a RLS.

Pomocí vytvořeného programu byly pořízeny nahrávky řeči a hluku vyskytujícího se zejména v okolním prostředí. Tyto nahrávky byly spolu sečteny pro získání zašuměné nahrávky a dále odfiltrovány adaptivními filtry. Minimální hodnota odstupu signálu od šumu, pro dodržení správné detekce řeči, byla vyhodnocena na hodnotu 4 dB. Pod touto hranicí SNR už detektor není schopen rozpoznat, zda se jedná o řeč či šum. U výpočtů globálního SNR se použitím filtrace pomocí LMS algoritmu hodnoty SNR pohybovali kolem 18 dB při správném nastavení algoritmu, zatímco při použití algoritmu RLS vyšly hodnoty SNR přes 30 dB, což zajištuje dostatečný poměr SNR i pro telekomunikační techniku. Při výpočtech segmentálního SNR, tedy v řečově aktivních segmentech, byly hodnoty při správném nastavení algoritmu o 10 dB vyšší pro oba použité algoritmy. Vyplývá z toho, že algoritmus LMS je sice jednoduchý a matematicky nenáročný, ale má pomalou rychlost konvergence a větší chybu filtrace. Oproti tomuto je RLS algoritmus matematicky složitý, ale výsledky ukázaly, že je velice přesný a rychlý. Adaptivní filtr využívající RLS algoritmus vykazuje lepší filtrační vlastnosti, ale na druhou stranu je náročný na výpočty.

Jako algoritmus pro rozpoznávání slov byla zvolena metoda dynamického borcení časové osy DTW. Metoda DTW je již starou metodou, ale svojí jednoduchostí a efektivitou je i dnes hojně využívána. Pro posuzování podobnosti pomocí této metody, je důležitý detektor řeči a jeho správné nastavení, aby nedocházelo k vyhodnocování řeči, slov, jak je zobrazeno na obrázku (obr. 23), tj. zachytávání okluzív. V posledním případě byly rozpoznávány referenční nahrávky s nahrávkami, ze kterých byl odfiltrován hluk křižovatky pomocí algoritmů LMS a RLS, a tím se zkoumala úspěšnost adaptivní filtrace. Podobnost se pohybovala pro oba algoritmy od 70 % do 95 %, tedy dostatečná podobnost pro rozpoznání slov z použitého slovníku. Rozdílem u použitých algoritmů byl výpočet odstupu signálu od šumu, kdy pro algoritmus RLS se hodnoty pohybovaly v rozmezí od 36 dB do 42 dB, zatímco u LMS algoritmu se pohybovaly od 20 dB 29 dB.

5 Seznam použité literatury

[1] KONDO, K., *Subjective Quality Measurement of Speech Its Evaluation, Estimation and Application*. Hardcover. ISBN 978-3-642-27505-0. [online]. 2012. Dostupné z: http://www.springer.com/gp/book/9783642275050

[2] VASEGHI, S., V. *Advanced Digital Signal Processing and Noise Reduction.* Fourth Edition, Ltd. ISBN: 978-0-470-75406-1 [online]. 2008. Dostupné z: http://sharif.ir/~bahram/sp4cl/MainReferences/advancedDigitalSignalProcessingAndNoiseReduction.pdf

[3] MARTINEK, Radek, ŽÍDEK, Jan. *Use of adaptive filtering for noise reduction in communications systems*. 2010 International Conference on Applied Electronics (AE) 2010. Plzeň: Západočeská univerzita, 2010. ISBN 978-80-7043-865-7, ISSN 1803-7232.

[4] PSUTKA, Josef. *Komunikace s počítačem mluvenou řečí*. 1. vyd., Praha: Academia, 1995. ISBN 80-200-0203-0

[5] ČERNOCKÝ, Jan. *Zpracování řečových signálů – studijní opora*. Vysoké učení technické v Brně [online]. 2006. Dostupné z: http://www.fit.vutbr.cz/study/courses/ZRE/public/opora/zre_opora.pdf

[6] ITU-T Rec.P.800, Series P: *Telephone transmission quality. Methods for subjective determination of transmission quality*, 1996.

[7] ITU-T Rec. P.862, Series P: *Telephone transmission quality, telephone installations, local line networks. Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band tele-phone networks and speech codecs*, 2001.

[8] ITU – T Rec. P.830, Series P: *Subjectives performance asessment of telephone-band and wideband digital codecs,* 1996.

[9] AKSAMÍT, J. *Metody subjektivního hodnocení kvality hovorových signálů* [online]. 2007.
Dostupné z http://access.feld.cvut.cz/view.php?cisloclanku=2007030002

[10] KOVAŘÍK, Jiří. *Objektivní metody hodnocení kvality řeči* [online]. Vysoké učení technické v Brně. Fakulta elektrotechniky a komunikačních technologií, 2010. Dostupné z: http://hdl.handle.net/11012/5769

[11] PSUTKA, Josef, MÜLLER, Luděk, MATOUŠEK, Jindřich, RADOVÁ, Vlasta. *Mluvíme s počítačem česky*. První vydání. Praha: Academia, 2006. ISBN 80-200-1309-1.

[12] PRASAD, R., SANGWAN, A., JAMADAGNI, H. *Comparison of Voice Activity Detection Algorithms for VoIP* [online]. 2002. Dostupné z: http://homepage.tudelft.nl/w5p50/pdffiles/Comparison%20of%20Voice%20Activity%20Detection%20Algorithms%20for%20VoIP.pdf

[13] KROČIL, Josef. *Interaktivní systém pro detekci řečového signálu* [online]. Ostrava, 2015. Diplomová práce, Vysoká škola báňská – Technická univerzita Ostrava. 2015.Dostupné z: http://dspace.vsb.cz/handle/10084/108574

[14] POLLÁK, P.*: Metody odhadu odstupu signálu od šumu v řečovém signálu*. Akustické listy, č. 7, 2001

[15] SOVKA, P. a POLLÁK, P. *Vybrané metody číslicového zpracování signálů* (in Czech). Vydavatelství ČVUT, Praha 6, 2003.
[16] PORUBA, J., MATĚJÍČEK, L., *Odfiltrování rušivých signálů ze zašumělé řeči.* [online] Dostupné z: http://www.elektrorevue.cz/clanky/02047/index.html

A. Přílohy na CD

I. Elektronická verze bakalářské práce
II. Programy
Vyhodnoceni kvality reci
Nahravani a ukladani dat 1 MIC
Nahravani a ukladani dat 2 MIC
Programy Matlab
III. Nahrávky použitých šumů a řeči
IV. Knihovny LabVIEW
Adapative filter toolkit
NI-DAQ